

# Finite Elements

Guido Kanschat

May 7, 2019

# Contents

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Elliptic PDE and Their Weak Formulation</b>               | <b>3</b>  |
| 1.1      | Elliptic boundary value problems . . . . .                   | 3         |
| 1.1.1    | Linear second order PDE . . . . .                            | 3         |
| 1.1.2    | Variational principle and weak formulation . . . . .         | 8         |
| 1.1.3    | Boundary conditions in weak form . . . . .                   | 10        |
| 1.2      | Hilbert Spaces and Bilinear Forms . . . . .                  | 11        |
| 1.3      | Fast Facts on Sobolev Spaces . . . . .                       | 20        |
| 1.4      | Regularity of Weak Solutions . . . . .                       | 24        |
| <b>2</b> | <b>Conforming Finite Element Methods</b>                     | <b>27</b> |
| 2.1      | Meshes, shape functions, and degrees of freedom . . . . .    | 27        |
| 2.1.1    | Shape function spaces on simplices . . . . .                 | 30        |
| 2.1.2    | Shape functions on tensor product cells . . . . .            | 31        |
| 2.1.3    | The Galerkin equations and Céa's lemma . . . . .             | 35        |
| 2.1.4    | Mapped finite elements . . . . .                             | 37        |
| 2.2      | A priori error analysis . . . . .                            | 40        |
| 2.2.1    | Approximation of Sobolev spaces by finite elements . . . . . | 41        |
| 2.2.2    | Estimates of stronger norms . . . . .                        | 45        |
| 2.2.3    | Estimates of weaker norms and linear functionals . . . . .   | 46        |
| 2.2.4    | Green's function and maximum norm estimates . . . . .        | 48        |
| 2.3      | A posteriori error analysis . . . . .                        | 50        |
| 2.3.1    | Quasi-interpolation in $H^1$ . . . . .                       | 50        |

|          |  |           |
|----------|--|-----------|
| <b>3</b> | <b>Variational Crimes</b>                              | <b>59</b> |
| 3.1      | Numerical quadrature . . . . .                         | 59        |
| <b>4</b> | <b>Solving the Discrete Problem</b>                    | <b>67</b> |
| 4.1      | The Richardson iteration . . . . .                     | 70        |
| 4.2      | The conjugate gradient method . . . . .                | 75        |
| 4.3      | Condition numbers of finite element matrices . . . . . | 79        |
| 4.4      | Multigrid methods . . . . .                            | 82        |
| <b>5</b> | <b>Discontinuous Galerkin methods</b>                  | <b>84</b> |
| 5.1      | Nitsche's method . . . . .                             | 84        |
| 5.2      | The interior penalty method . . . . .                  | 89        |
| 5.2.1    | Bounded formulation in $H^1$ . . . . .                 | 93        |

# Chapter 1

## Elliptic PDE and Their Weak Formulation

### 1.1 Elliptic boundary value problems

#### 1.1.1 Linear second order PDE

**1.1.1 Notation:** Dimension of “physical space” will be denoted by  $d$ . We denote coordinates in  $\mathbb{R}^d$  as

$$\mathbf{x} = (x_1, \dots, x_d)^T.$$

In the special cases  $d = 2, 3$  we also write

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix},$$

respectively. The Euclidean norm on  $\mathbb{R}^d$  is denoted as

$$|\mathbf{x}| = \sqrt{\sum_{i=1}^d x_i^2}.$$

**1.1.2 Notation:** Partial derivatives of a function  $u \in C^1(\mathbb{R}^d)$  are denoted by

$$\frac{\partial u(\mathbf{x})}{\partial x_i} = \frac{\partial}{\partial x_i} u(\mathbf{x}) = \partial_{x_i} u(\mathbf{x}) = \partial_i u(\mathbf{x}).$$

The **gradient** of  $u \in C^1$  is the row vector

$$\nabla u = (\partial_1 u, \dots, \partial_d u)$$

The **Laplacian** of a function  $u \in C^2(\mathbb{R}^d)$  is

$$\Delta u = \partial_1^2 u + \dots + \partial_d^2 u = \sum_{i=1}^d \partial_i^2 u$$

**1.1.3 Notation:** When we write equations, we typically omit the independent variable  $\mathbf{x}$ . Therefore,

$$\Delta u \equiv \Delta u(\mathbf{x}).$$

**1.1.4 Definition:** A linear PDE of second order in divergence form for a function  $u \in C^2(\mathbb{R}^d)$  is an equation of the form

$$-\sum_{i,j=1}^d \partial_i(a_{ij}(\mathbf{x})\partial_j u) + \sum_{i=1}^d (b_i(\mathbf{x})\partial_i u) + c(\mathbf{x})u = f(x) \quad (1.1)$$

**1.1.5 Definition:** An important model problem for the equations we are going to study is **Poisson's equation**

$$-\Delta u = f. \quad (1.2)$$

**1.1.6.** Already with ordinary differential equations we experience that we typically do not search for solutions of the equation itself, but that we “anchor” the solution by solving an initial value problem, fixing the solution at one point on the time axis.

It does not make sense to speak about an initial point in  $\mathbb{R}^d$ . Instead, it turns out that it is appropriate to consider solutions on certain subsets of  $\mathbb{R}^d$  and impose conditions at the boundary.

**1.1.7 Definition:** A **domain** in  $\mathbb{R}^d$  is a connected, open set of  $\mathbb{R}^d$ . We typically use the notation  $\Omega \subset \mathbb{R}^d$ .

The **boundary** of a domain  $\Omega$  is denoted by  $\partial\Omega$ . To any point  $\mathbf{x} \in \partial\Omega$ , we associate the outer unit **normal vector**  $\mathbf{n} \equiv \mathbf{n}(\mathbf{x})$ .

The symbol  $\partial_n u \equiv (\nabla u) \cdot \mathbf{n}$  denotes the **normal derivative** of a function  $u \in C^1(\overline{\Omega})$  at a point  $\mathbf{x} \in \partial\Omega$ .

**1.1.8 Definition:** We distinguish three types of boundary conditions for Poisson's equation, namely for a point  $\mathbf{x} \in \partial\Omega$  with a given function  $g$

1. Dirichlet:

$$u(\mathbf{x}) = g(\mathbf{x})$$

2. Neumann:

$$\partial_n u(\mathbf{x}) = g(\mathbf{x})$$

3. Robin: for some positive function  $\alpha$  on  $\partial\Omega$

$$\partial_n u(\mathbf{x}) + \alpha(\mathbf{x})u(\mathbf{x}) = g(\mathbf{x})$$

While only one of these boundary conditions can hold in a single point  $\mathbf{x}$ , different boundary conditions can be active on different subsets of  $\partial\Omega$ . We denote such subsets as  $\Gamma_D$ ,  $\Gamma_N$ , and  $\Gamma_R$ .

**1.1.9 Definition:** The **Dirichlet problem** for Poisson's equation (in differential form) is: find  $u \in C^2(\Omega) \cap C(\overline{\Omega})$ , such that

$$-\Delta u(\mathbf{x}) = f(\mathbf{x}) \quad x \in \Omega, \quad (1.3a)$$

$$u(\mathbf{x}) = g(\mathbf{x}) \quad x \in \partial\Omega. \quad (1.3b)$$

Here, the functions  $f$  on  $\Omega$  and  $g$  on  $\partial\Omega$  are data of the problem. The Dirichlet problem is called **homogeneous**, if  $g \equiv 0$ .

**1.1.10 Theorem (Dirichlet principle):** If a function  $u \in C^2(\Omega) \cap C(\overline{\Omega})$  solves the Dirichlet problem, then it minimizes the **Dirichlet energy**

$$E(v) = \int_{\Omega} \frac{1}{2} |\nabla v|^2 \, d\mathbf{x} - \int_{\Omega} f v \, d\mathbf{x}, \quad (1.4)$$

among all functions  $v$  from the set

$$V_g = \{v \in C^2(\Omega) \cap C(\overline{\Omega}) \mid v|_{\partial\Omega} = g\}. \quad (1.5)$$

This minimizer is unique.

*Proof.* Using variation of  $E$ , we will show that

$$\left. \frac{d}{d\varepsilon} E(u + \varepsilon v) \right|_{\varepsilon=0} = 0$$

for all  $v \in V_0$  since this implies  $u + \varepsilon v = g$  on  $\partial\Omega$ . By evaluating the square we have

$$\frac{d}{d\varepsilon} E(u + \varepsilon v) = \int_{\Omega} \nabla u \nabla v + \varepsilon |\nabla v|^2 - f v \, dx$$

Since we are interested in  $E(u)$ , we now consider  $\varepsilon = 0$ . We get that  $u$  minimizes  $E(u + \varepsilon v)$  at  $\varepsilon = 0$  implies

$$\int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} f v \, dx, \quad \forall v \in V_0.$$

By Green's formula

$$\int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} -\Delta u v \, dx + \int_{\partial\Omega} \partial_n u v \, ds$$

we obtain that if  $u$  minimizes  $E(\cdot)$ , then

$$\int_{\Omega} \nabla u \nabla v \, dx = \int_{\Omega} f v \, dx, \quad \forall v \in V_0,$$

since  $v \in V_0$  vanishes on  $\partial\Omega$ . In summary, we have proven so far that if  $u$  solves Poisson's Equation, then it is a stationary point of  $E(\cdot)$ . It remains to show that  $E(u) \leq E(u + v)$  for any  $v \in V_0$ . Using  $\int_{\Omega} f v \, dx = \int_{\Omega} \nabla u \nabla v$  yields

$$E(u+v) - E(u) = \frac{1}{2} \int_{\Omega} |\nabla(u+v)|^2 - 2\nabla(u+v) \nabla u + |\nabla u|^2 \, dx = \frac{1}{2} \int_{\Omega} |\nabla v|^2 \, dx \geq 0.$$

This also proves uniqueness.  $\square$

**1.1.11 Lemma:** A minimizing sequence for the Dirichlet energy exists and it is a Cauchy sequence.

*Proof.* The Dirichlet energy  $E(\cdot)$  is bounded from below and hence an infimum exists. Thus, there also exists a series  $\{u^{(n)}\}_{n \in \mathbb{N}}$  converging to this infimum, i.e.

$$\lim_{n \rightarrow \infty} E(u^{(n)}) = \inf_{v \in V_0} E(v).$$

Second, we show that  $\{u^{(n)}\}_n$  is a Cauchy sequence.

For the first part we use Friedrich's inequality

$$\|v\|_{L^2(\Omega)} \leq \lambda(\Omega) \|\nabla v\|_{L^2(\Omega)} \quad v \in V_0.$$

The proof of this result will be given later. Using Hölder's inequality we obtain

$$E(v) = \frac{1}{2} \|\nabla v\|_{L^2(\Omega)}^2 - \int_{\Omega} f v \, dx \geq \frac{1}{2} \|\nabla v\|_{L^2(\Omega)}^2 - \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)}$$

Applying Friedrich's inequality yields that the above expression is greater or equal than

$$\frac{1}{2} \|\nabla v\|_{L^2(\Omega)}^2 - \|\nabla v\|_{L^2(\Omega)} \frac{1}{\lambda(\Omega)} \|f\|_{L^2(\Omega)}.$$

Finally, we apply Young's inequality  $ab \leq 1/2(a^2 + b^2)$  to obtain

$$\frac{1}{2} \|\nabla v\|_{L^2(\Omega)}^2 - \|\nabla v\|_{L^2(\Omega)} \frac{1}{2\lambda(\Omega)} \|f\|_{L^2(\Omega)}^2$$

which yields  $E(v) \geq -\frac{1}{2\lambda(\Omega)^2} \|f\|_{L^2(\Omega)}^2$  as a lower bound independent of  $v$ . To prove the second part, we use the parallelogram identity  $|v + w|^2 + |v - w|^2 = 2|v|^2 + 2|w|^2$ . Let  $m, n$  be natural numbers, then

$$\begin{aligned} |u^{(n)} - u^{(m)}|_1^2 &= 2|u^{(n)}|_1^2 + 2|u^{(m)}|_1^2 - 4|1/2(u^{(n)} + u^{(m)})|_1^2 \\ &= 4E(u^{(n)}) + 4 \int_{\Omega} f u^{(n)} \, dx + 4E(u^{(m)}) + 4 \int_{\Omega} f u^{(m)} \, dx \\ &\quad - 8E(1/2(u^{(n)} + u^{(m)})) - 8 \int_{\Omega} 1/2 f (u^{(n)} + u^{(m)}) \\ &= 4E(u^{(n)}) + 4E(u^{(m)}) - 8E(1/2(u^{(n)} + u^{(m)})) \end{aligned}$$

Taking the limit  $m, n \rightarrow \infty$  yields  $4E(u^{(n)}) + 4E(u^{(m)}) \rightarrow 8 \inf_{v \in V_0} E(v)$ . Lastly,  $-E(1/2(u^{(n)} + u^{(m)}))$  can be bounded by  $\inf_{v \in V_0} E(v)$ . It follows that  $\limsup_{m, n \rightarrow \infty} |u^{(n)} - u^{(m)}|_1^2 \leq 0$  and consequently as desired

$$\lim_{m, n \rightarrow \infty} |u^{(n)} - u^{(m)}|_1^2 = 0.$$

□



**Note 1.1.12.** Dirichlet-proof Dirichlet's principle proved essential for the development of a rigorous solution theory for Poisson's equation. Its proof will be deferred to the next theorem.

### 1.1.2 Variational principle and weak formulation

**1.1.13 Theorem:** A function  $u \in V_g$  minimizes the Dirichlet energy, if and only if there holds

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx, \quad \forall v \in V_0. \quad (1.6)$$

Moreover, any solution to the Dirichlet problem in Definition 1.1.9 solves this equation.

**1.1.14 Corollary:** If a minimizer of the Dirichlet energy exists, it is necessarily unique.

**1.1.15 Lemma:** A function  $u \in V_g$  minimizes the Dirichlet energy admits the representation  $u = u_g + u_0$ , where  $u_g \in V_g$  is arbitrary and  $u_0 \in V_0$  solves

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx - \int_{\Omega} \nabla u_g \cdot \nabla v \, dx, \quad \forall v \in V_0. \quad (1.7)$$

The function  $u_0$  depends on the choice of  $u_g$ , but not the minimizer  $u$ .

**1.1.16 Notation:** The inner product of  $L^2(\Omega)$  is denoted by

$$(u, v) \equiv (u, v)_{\Omega} \equiv (u, v)_{L^2(\Omega)} = \int_{\Omega} uv \, d\mathbf{x}.$$

Its norm is

$$\|u\| \equiv \|u\|_{\Omega} \equiv \|u\|_{L^2(\Omega)} \equiv \|u\|_{L^2} = \sqrt{(u, v)_{L^2(\Omega)}}.$$

**1.1.17 Lemma (Friedrichs inequality):** For any function in  $v \in V_0$  there holds

$$\|v\|_{\Omega} \leq \text{diam}(\Omega) \|\nabla v\|_{\Omega}. \quad (1.8)$$

**1.1.18 Lemma:** The definitions

$$\begin{aligned} |v|_1 &= \|\nabla v\|_{L^2(\Omega)}, \\ \|v\|_1 &= \sqrt{\|v\|_{L^2(\Omega)}^2 + |v|_1^2}, \end{aligned} \tag{1.9}$$

both define a norm on  $V_0$ .

**1.1.19 Problem:** Prove the Friedrichs inequality.

**1.1.20 Lemma:** The Dirichlet energy with homogeneous boundary conditions is bounded from below and thus has an infimum. In particular, there exists a **minimizing sequence**  $\{u^n\}$  such that as  $n \rightarrow \infty$ ,

$$E(u^n) \rightarrow \inf_{v \in V_0} E(v). \tag{1.10}$$

**1.1.21 Lemma:** The minimizing sequence for the Dirichlet energy is a Cauchy sequence.

**1.1.22 Definition:** The completion of  $V_0$  under the norm  $\|v\|_1$  is the **Sobolev space**  $H_0^1(\Omega)$ .

**1.1.23 Lemma (Friedrichs inequality):** For any function in  $v \in H_0^1$  there holds

$$\|v\|_\Omega \leq \text{diam}(\Omega) \|\nabla v\|_\Omega. \tag{1.11}$$

*Proof.* Let  $v \in H_0^1(\Omega)$ . We make use of the fact, that by definition of  $H_0^1(\Omega)$ , there is a sequence  $v_n \rightarrow v$  with  $v_n \in V_0$ . By Lemma 1.1.17, Friedrichs' inequality holds for  $v_n$  uniformly in  $n$ . We conclude

$$\begin{aligned} \|v\|_\Omega &\leq \|v - v_n\|_\Omega + \|v_n\|_\Omega \\ &\leq \|v - v_n\|_\Omega + \text{diam } \Omega \|\nabla v_n\|_\Omega \\ &\leq \|v - v_n\|_\Omega + \text{diam } \Omega (\|\nabla v_n - \nabla v\|_\Omega + \|\nabla v\|_\Omega) \end{aligned}$$

As  $n \rightarrow \infty$ , the norms of the differences converge to zero, such that the desired result holds in the limit.  $\square$

**1.1.24 Definition:** The Dirichlet problem for Poisson's equation in weak form reads: find  $u \in H_g^1(\Omega)$  such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx, \quad \forall v \in H_0^1(\Omega). \quad (1.12)$$

**1.1.25 Theorem:** The weak formulation in Definition 1.1.24 has a unique solution.

### 1.1.3 Boundary conditions in weak form

**1.1.26 Lemma:** Let  $u \in V = H^1(\Omega)$  be a solution to the weak formulation

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx, \quad \forall v \in V(\Omega). \quad (1.13)$$

If  $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$  and  $\Omega$  has  $C^1$ -boundary, then  $u$  solves the boundary value problem

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ \partial_n u &= 0 && \text{on } \partial\Omega. \end{aligned} \quad (1.14)$$

**1.1.27 Definition:** A boundary condition inherent in the weak formulation and not explicitly stated is called **natural boundary condition**. If boundary values are obtained by constraining the function space it is called **essential boundary condition**.

We also call a boundary condition in strong form, if it is a constraint on the function space, and in weak form, if it is part of the weak formulation.

**Remark 1.1.28.** Dirichlet and homogeneous Neumann boundary conditions are examples for essential and natural boundary conditions, respectively.

**1.1.29 Lemma:** The boundary value problem

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{on } \Gamma_D \subset \partial\Omega \\ \partial_n u + \alpha u &= g && \text{on } \Gamma_R \subset \partial\Omega, \end{aligned} \quad (1.15)$$

has the weak form: find  $u \in V$  such that

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\Gamma_R} \alpha uv \, ds = \int_{\Omega} f v \, dx + \int_{\Gamma_R} g v \, ds, \quad \forall v \in V(\Omega). \quad (1.16)$$

## 1.2 Hilbert Spaces and Bilinear Forms

**1.2.1 Definition:** Let  $V$  be a vector space over  $\mathbb{R}$ . An **inner product** on  $V$  is a mapping  $\langle \cdot, \cdot \rangle : V \times V \rightarrow \mathbb{R}$  with the properties

$$\langle \alpha x + y, z \rangle = \alpha \langle x, z \rangle + \langle y, z \rangle \quad \forall x, y, z \in V; \alpha \in \mathbb{K} \quad (1.17)$$

$$\langle x, y \rangle = \langle y, x \rangle \quad \forall x, y \in V \quad (1.18)$$

$$\langle x, x \rangle \geq 0 \quad \forall x \in V \quad \text{and} \quad (1.19)$$

$$\langle x, x \rangle = 0 \Leftrightarrow x = 0, \quad (1.20)$$

usually referred to as (bi-)linearity, symmetry, and positive definiteness. We note that linearity in the second argument follows immediately by symmetry.

**1.2.2 Theorem (Bunyakovsky-Cauchy-Schwarz inequality):** For every inner product there holds the inequality

$$\langle v, w \rangle \leq \sqrt{\langle v, v \rangle} \sqrt{\langle w, w \rangle}. \quad (1.21)$$

Equality holds if and only if  $v$  and  $w$  are collinear.

*Proof.* The case  $w = 0$  is trivial. Without loss of generality we can therefore assume that  $w \neq 0$ . Define  $\lambda \in \mathbb{R}$  as  $\lambda = \frac{\langle v, w \rangle}{\langle w, w \rangle}$ . By (1.19) we have

$$0 \leq \langle v - \lambda w, v - \lambda w \rangle$$

and by (1.18) the right-hand side extends to

$$\begin{aligned} & \langle v, v \rangle - \langle v, \lambda w \rangle - \langle \lambda w, v \rangle + \langle \lambda w, \lambda w \rangle \\ &= \langle v, v \rangle - \bar{\lambda} \langle v, w \rangle - \bar{\lambda} \langle v, w \rangle + \lambda \bar{\lambda} \langle w, w \rangle. \end{aligned}$$

Evaluating  $\lambda$  yields the inequality

$$0 \leq \langle v, v \rangle - \frac{\overline{\langle v, w \rangle} \langle v, w \rangle}{\langle w, w \rangle} - \frac{\overline{\langle v, w \rangle} \langle v, w \rangle}{\langle w, w \rangle} + \frac{\overline{\langle v, w \rangle} \langle v, w \rangle \langle w, w \rangle}{\langle w, w \rangle^2}.$$

The result follows from multiplication with  $\langle w, w \rangle$  and arranging the summands.

For the second part let  $v, w$  be colinear, i.e. there is a  $\lambda \in \mathbb{K}$  such that  $v = \lambda w$ . Then deducing the equality is trivial. Now let equality hold for (1.21). We immediately get that the equality must also hold for

$$0 = \langle v - \lambda w, v - \lambda w \rangle.$$

However, by (1.20) this implies

$$0 = v - \lambda w.$$

Thus,  $v$  and  $w$  are colinear. □

**1.2.3 Lemma:** Every inner product defines a norm by

$$\|v\| = \sqrt{\langle v, v \rangle}. \quad (1.22)$$

*Proof.* Definiteness and homogeneity follow from the properties of the inner product. It remains to show the triangle inequality

$$\|u + v\| \leq \|u\| + \|v\|.$$

Squaring the left hand side yields with the Bunyakovsky-Cauchy-Schwarz inequality

$$\begin{aligned} \|u + v\|^2 &= \langle u + v, u + v \rangle = \|u\|^2 + 2\langle u, v \rangle + \|v\|^2 \\ &\leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 = (\|u\| + \|v\|)^2. \end{aligned}$$

□

**1.2.4 Definition:** A space  $V$  with is **complete** with respect to a norm, if all Cauchy sequences with elements in  $V$  have their limit in  $V$ . A subspace  $W \subset V$  is **closed** if it is complete in the topology of  $V$ . The **completion** of a space  $V$  with respect to a norm consists of the space  $V$  and the limits of all Cauchy sequences in  $V$ . We denote the completion of a space  $V$  by

$$\bar{V} = \bar{V}^{\|\cdot\|_V}. \quad (1.23)$$

**1.2.5 Definition:** A **normed vector space** is a vector space  $V$  with a norm  $\|\cdot\|$ . We may also write  $\|\cdot\|_V$  to highlight the connection. A normed vector space  $V$  which is complete with respect to its norm is called a **Banach space**. A vector space  $V$  equipped with an inner product  $\langle \cdot, \cdot \rangle$  is called an **inner product space** or **pre-Hilbert space**. A **Hilbert space** is a pre-Hilbert space which is also complete.

**1.2.6 Definition:** Let  $V$  be an inner product space over a field  $\mathbb{K}$ . Two vectors  $x, y \in V$  are called **orthogonal** if  $\langle x, y \rangle = 0$ . We write  $x \perp y$ . Let  $W$  be a subspace of  $V$ . We say that a vector  $v$  is orthogonal to the subspace  $W$ , if it is orthogonal to every vector in  $W$ . A set of nonzero mutually orthogonal vectors  $\{x_i\} \subset V$  is called **orthogonal set**. If additionally  $\|x_i\| = 1$  for all vectors, it is called an **orthonormal set**. These notions transfer directly from finite to countable sets.

**1.2.7 Definition:** Let  $W \subset V$  be a subspace of a Hilbert space  $V$ . We define its **orthogonal complement**  $W^\perp \subset V$  by

$$W^\perp = \{v \in V \mid \langle v, w \rangle_V = 0 \ \forall w \in W\}. \quad (1.24)$$

**1.2.8 Lemma:** The orthogonal complement  $W^\perp$  of a subspace  $W \subset V$  is closed in the sense of Definition 1.2.4.

*Proof.* By the Bunyakovsky-Cauchy-Schwarz inequality, the inner product is continuous on  $V \times V$ . Therefore, the mapping

$$\begin{aligned} \varphi_w: V &\rightarrow \mathbb{R}, \\ v &\mapsto \langle v, w \rangle, \end{aligned}$$

is continuous. For any  $w \in W$ , the kernel of  $\varphi_w$  is closed as the pre-image of the closed set  $\{0\}$ . Since

$$W^\perp = \bigcap_{w \in W} \ker(\varphi_w),$$

it is closed as the intersection of closed sets.  $\square$

**1.2.9 Theorem:** Let  $W$  be a subspace of a Hilbert space  $V$  and  $W^\perp$  its orthogonal complement. Then,  $W^\perp = \overline{W}^\perp$ . Further,  $V = W \oplus W^\perp$  if and only if  $W$  is closed.

*Proof.* Clearly,  $\overline{W}^\perp \subset W^\perp$  since  $W \subset \overline{W}$ . Let now  $u \in W^\perp$ . Then,  $\varphi = \langle u, \cdot \rangle$  is a continuous linear functional on  $V$ . Therefore, if a sequence  $w_n \subset W$  converges to  $w \in \overline{W}$ , we have

$$\langle u, w \rangle = \lim_{n \rightarrow \infty} \langle u, w_n \rangle = 0,$$

since  $u \in W^\perp$ . Hence,  $u \in \overline{W}^\perp$  and  $W^\perp = \overline{W}^\perp$ .

Now, the “only if” follows by the fact, that if  $W$  is not closed, there is an element  $w \in \overline{W}$  but not in  $W$  such that  $\langle w, u \rangle = 0$  for all  $u \in W^\perp$ . Thus,  $w \notin W^\perp$  and consequently  $w \notin W^\perp \oplus W$ .

Let now  $W$  be closed. We show that for all  $v \in V$  there is a unique decomposition

$$v = w + u, \quad \text{with} \quad w \in W, u \in W^\perp. \quad (1.25)$$

This is equivalent to  $V = W \oplus W^\perp$ . Uniqueness follows, since

$$v = w_1 + u_1 = w_2 + u_2$$

implies that for any  $y \in V$

$$0 = \langle w_1 - w_2 + u_1 - u_2, y \rangle = \langle w_1 - w_2, y \rangle + \langle u_1 - u_2, y \rangle.$$

Choosing  $y = u_1 - u_2$  and  $w_1 - w_2$  in turns, we see that one of the inner products vanishes for orthogonality and the other implies that the difference is zero.

If  $v \in W$ , we choose  $w = v$  and  $u = 0$ . For  $v \notin W$ , we prove existence by considering that due to the closedness of  $W$  there holds

$$d = \inf_{w' \in W} \|v - w'\| > 0.$$

Let  $w_n$  be a minimizing sequence. Using the parallelogram identity

$$\|a + b\|^2 + \|a - b\|^2 = 2\|a\|^2 + 2\|b\|^2,$$

we prove that  $\{w_n\}$  is a Cauchy sequence by

$$\begin{aligned}
\|w_m - w_n\|^2 &= \|(v - w_n) - (v - w_m)\|^2 \\
&= 2\|v - w_n\|^2 + 2\|v - w_m\|^2 - \|2v - w_m - w_n\|^2 \\
&= 2\|v - w_n\|^2 + 2\|v - w_m\|^2 - 4\left\|v - \frac{w_m + w_n}{2}\right\|^2 \\
&\leq 2\|v - w_n\|^2 + 2\|v - w_m\|^2 - 4d^2,
\end{aligned}$$

since  $(w_m + w_n)/2 \in W$  and  $d$  is the infimum. Now we use the minimizing property to obtain

$$\lim_{m,n \rightarrow \infty} \|w_m - w_n\|^2 = 2d^2 + 2d^2 - 4d^2 = 0.$$

Since  $V$  is given as a Hilbert space and as such complete,  $w = \lim w_n$  exists and by the closedness of  $W$ , we have  $w \in W$ . Let  $u = v - w$ . By continuity of the norm, we have  $\|u\| = d$ . It remains to show that  $u \in W^\perp$ . To this end, we introduce the variation  $w + \varepsilon \tilde{w}$  with  $\tilde{w} \in W$  to obtain

$$\begin{aligned}
d^2 &\leq \|v - w - \varepsilon \tilde{w}\|^2 \\
&= \|u\|^2 - 2\varepsilon \langle u, \tilde{w} \rangle + \varepsilon^2 \|\tilde{w}\|^2,
\end{aligned}$$

implying for any  $\varepsilon > 0$

$$0 \leq -2\varepsilon \langle u, \tilde{w} \rangle + \varepsilon^2 \|\tilde{w}\|^2,$$

which requires  $\langle u, \tilde{w} \rangle = 0$ . Since  $\tilde{w} \in W$  was chosen arbitrarily, we have  $u \in W^\perp$ .  $\square$

**1.2.10 Definition:** Let  $W$  be a closed subspace of the Hilbert space  $V$  and  $W^\perp$  be its orthogonal complement. Then, the **orthogonal projection** operators

$$\begin{aligned}
\Pi_W &: V \rightarrow W \\
\Pi_{W^\perp} &: V \rightarrow W^\perp
\end{aligned} \tag{1.26}$$

are defined by the unique decomposition

$$v = \Pi_W v + \Pi_{W^\perp} v. \tag{1.27}$$



**1.2.11 Definition:** A **linear functional** on a vector space  $V$  is a linear mapping from  $V$  to  $\mathbb{K}$ .

The **dual space**  $V^*$  of a vector space  $V$ , also called the **normed dual**, is the space of all bounded linear functionals on  $V$  equipped with the norm

$$\|\varphi\|_{V^*} = \sup_{v \in V} \frac{\varphi(v)}{\|v\|_V}. \quad (1.28)$$

**1.2.12 Theorem (Riesz representation theorem):** Let  $V$  be a Hilbert space. Then,  $V$  is isometrically isomorphic to  $V^*$ . In particular, there is an isomorphism

$$\begin{aligned} \varrho: V &\rightarrow V^*, \\ y &\mapsto f, \end{aligned} \quad (1.29)$$

such that

$$\begin{aligned} \langle x, y \rangle &= f(x) \quad \forall x \in V, \\ \|y\|_V &= \|f\|_{V^*}. \end{aligned} \quad (1.30)$$

We refer to  $\varrho$  as **Riesz isomorphism**.

*Proof.* The proof is constructive and makes use of the orthogonal complement.

First, it is clear that for any  $y \in V$  a linear functional  $f \in V^*$  is defined by  $f(\cdot) = \langle \cdot, y \rangle$ . Furthermore,  $\varrho$  is injective, since

$$\langle x, y \rangle = 0 \quad \forall x \in V$$

implies  $y \in V^\perp = \{0\}$ . By the Bunyakovsky-Cauchy-Schwarz inequality, we have

$$\|f\|_{V^*} = \sup_{x \in V} \frac{|f(x)|}{\|x\|_V} = \sup_{x \in V} \frac{|\langle x, y \rangle|}{\|x\|_V} \leq \|y\|_V,$$

with equality for  $x = y$ . It remains to show that  $\varrho$  is surjective. To this end, let  $f \in V^*$  be arbitrary and let  $N = \ker(f)$ . If  $N = V$ , we choose  $y = 0$ . If not, choose  $y^\perp \in N^\perp$  and let

$$y = \frac{f(y^\perp)}{\|y^\perp\|^2} y^\perp \in N^\perp, \quad (1.31)$$

such that  $f(y) = |f(y^\perp)|^2 / \|y^\perp\|^2 \neq 0$ . Let now  $x \in V$  be chosen arbitrarily. Then, there holds

$$x = \left( x - \frac{f(x)}{f(y)} y \right) + \frac{f(x)}{f(y)} y,$$

where  $\frac{f(x)}{f(y)}$  denotes a scalar. Since

$$f \left( x - \frac{f(x)}{f(y)} y \right) = \left( f(x) - f(x) \frac{f(y)}{f(y)} \right) = 0,$$

this decomposition amounts to  $x = x^0 + x^\perp$  with  $x^0 \in N$  and  $x^\perp \in N^\perp$ . It is unique according to Definition 1.2.10. Thus, we have that  $x^\perp$  is a multiple of  $y$ , say  $x^\perp = \alpha y$  with  $\alpha = \frac{f(x)}{f(y)}$  and thus

$$\begin{aligned} f(x) &= f(x^0) + f(x^\perp) &= \alpha f(y) &= \alpha \frac{|f(y^\perp)|^2}{\|y^\perp\|^2} \\ \langle x, y \rangle &= \langle x^0, y \rangle + \langle x^\perp, y \rangle &= \alpha \|y\|_V^2 &= \alpha \frac{|f(y^\perp)|^2}{\|y^\perp\|^4} \|y^\perp\|^2 \end{aligned}$$

Hence, the two terms are equal and  $\varrho$  is surjective.  $\square$

**1.2.13 Definition:** A **bilinear form**  $a(.,.)$  on a Hilbert space  $V$  is a mapping  $a: V \times V \rightarrow \mathbb{R}$ , which is linear in both arguments. The bilinear form is **bounded**, if there is a constant  $M$  such that

$$a(u, v) \leq M \|u\|_V \|v\|_V, \quad \forall u, v \in V. \quad (1.32)$$

It is called **coercive** or **elliptic**, if there is a constant  $\alpha$  such that

$$a(u, u) \geq \alpha \|u\|_V^2 \quad \forall u \in V. \quad (1.33)$$

**Notation 1.2.14.** One often speaks of  **$V$ -elliptic** instead of elliptic to point out that the ellipticity of the bilinear form indeed depends on the scalarproduct inducing the  $V$ -norm.

**1.2.15 Lemma:** Let  $a_{ij} \in C^1(\Omega)$ ,  $b_i, c \in C^0(\Omega)$ . A solution to the Dirichlet problem

$$\begin{aligned} - \sum_{i,j=1}^d \partial_i(a_{ij} \partial_j u) + \sum_{i=1}^d (b_i \partial_i u) + cu &= f && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega \end{aligned} \quad (1.34)$$

solves the weak problem: find  $u \in V_0$  such that for all  $v \in V_0$

$$a(u, v) \equiv (\mathbf{A} \nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) + (cu, v) = (f, v), \quad (1.35)$$

where  $\mathbf{A}(\mathbf{x}) = (a_{ij}(\mathbf{x}))$  is the matrix of coefficients of the second order term and  $\mathbf{b}(\mathbf{x}) = (b_i(\mathbf{x}))$  is the vector of coefficients of the first order term.

If additionally  $u \in C^2(\Omega)$  holds, then the solution to the weak problem solves the Dirichlet problem in differential form.

**1.2.16 Lemma (Lax-Milgram):** Let  $a(.,.)$  be a bounded, coercive bilinear form on a Hilbert space  $V$  and let  $f \in V^*$ . Then, there is a unique element  $u \in V$  such that

$$a(u, v) = f(v) \quad \forall v \in V. \quad (1.36)$$

Furthermore, there holds

$$\|u\|_V \leq \frac{1}{\alpha} \|f\|_{V^*}. \quad (1.37)$$

*Proof.* To prove Lax-Milgram we first consider uniqueness and then the existence of a solution.

Assume that there are solutions  $u_1, u_2 \in V$  of (1.36), i.e. there holds  $a(u_1, v) = f(v)$  and  $a(u_2, v) = f(v)$  for all  $v \in V$ . Thus,  $a(u_1 - u_2, v) = 0$  for all  $v \in V$ . Now choose  $v = u_1 - u_2 \in V$ . Since  $a(.,.)$  is coercive with  $\alpha > 0$  it holds

$$0 = a(u_1 - u_2, u_1 - u_2) \geq \alpha \|u_1 - u_2\|_V^2$$

which implies  $u_1 - u_2 = 0$ . Hence  $u_1 = u_2$ .

Let us now consider the existence of a solution. We will define a linear functional to apply Riesz and Banach fixed point theorem. For all  $y \in V$  there holds

$$\langle y, \cdot \rangle - \omega [a(y, \cdot) - f(\cdot)] \in V^*$$

with  $\omega > 0$ . Due to Riesz there exists an isomorphism  $\varrho : V \rightarrow V^*$  that maps a given  $z \in V$  to  $\langle y, \cdot \rangle - \omega [a(y, \cdot) - f(\cdot)]$  such that

$$\langle v, z \rangle = \langle y, v \rangle - \omega [a(y, v) - f(v)] \quad \forall v \in V.$$

Now we define the mapping  $T_\omega : V \rightarrow V$  that maps  $y \mapsto z$  and define  $S : V \rightarrow V$  such that

$$\langle Su, v \rangle = a(u, v) \quad \forall v \in V$$

with  $\|Su\|_V \leq M\|u\|_V$  for  $M > 0$ . Now consider  $y - x$  instead of  $y$ . This leads for all  $v \in V$  to

$$\langle T_\omega y - T_\omega x, v \rangle = \langle y - x, v \rangle - \omega [a(y - x, v)] = \langle y - x, v \rangle - \omega \langle S(y - x), v \rangle.$$

Thus, we can conclude that  $T_\omega(y - x) = y - x - \omega S(y - x)$ . Applying the norm and using the fact that it is induced by the inner product of  $V$  we get

$$\begin{aligned} \|T_\omega y - T_\omega x\|_V^2 &= \|y - x\|_V^2 - 2\omega \langle S(y - x), y - x \rangle + \omega^2 \|S(y - x)\|^2 \\ &\leq \|y - x\|^2 - 2\omega a(y - x, y - x) + \omega^2 M^2 \|y - x\|^2 \\ &\leq (1 - 2\omega\alpha + \omega^2 M^2) \|y - x\|^2. \end{aligned}$$

As we want to apply Banach fixed point theorem, we need  $T_\omega$  to be a contraction. Therefore, we need  $1 - 2\omega\alpha + \omega^2 M^2 < 1$  which is given for  $\omega \in (0, \frac{2\alpha}{M^2})$ . Hence,  $T_\omega$  is a contraction and there exists a  $\nu \in V$  such that  $T_\omega \nu = \nu$  and  $\langle T_\omega \nu, \nu \rangle = \langle \nu, \nu \rangle - \omega [a(\nu, \nu) - f(\nu)]$  which implies  $a(\nu, \nu) = f(\nu)$ . Thus, there exists a solution  $\nu$ .

Now the stability estimate (1.37) is left to prove. Using the coercivity of our bilinear form yields

$$\alpha \|u\|_V \leq \frac{a(u, u)}{\|u\|_V} = \frac{f(u)}{\|u\|_V} \leq \sup_{u \in V} \frac{|f(u)|}{\|u\|_V} = \|f\|_{V^*}.$$

□

**1.2.17 Lemma:** Let  $a_{ij}, c \in L^\infty(\Omega)$ ,  $b_i \in C^1(\overline{\Omega})$  such that there holds for a positive constant  $\alpha$

$$\begin{aligned} \alpha |\xi|^2 &\leq \xi^T \mathbf{A}(\mathbf{x}) \xi, \quad \forall \xi \in \mathbb{R}^d, \\ 0 &\leq c + \nabla \cdot \mathbf{b}. \end{aligned} \tag{1.38}$$

Then, the associated bilinear form is coercive and bounded on  $H_0^1(\Omega)$ , and thus the weak formulation has a unique solution.

**1.2.18 Definition:** A differential equation of second order in divergence form (1.34) such that (1.38) holds is called **elliptic**. The lower bound  $\alpha$  is the ellipticity constant.

## 1.3 Fast Facts on Sobolev Spaces

**1.3.1.** While this section reviews some of the basic mathematical properties of Sobolev spaces, it suffers a bit from overly abstract mathematical arguments. While the strongly mathematically inclined reader might appreciate this, it is not really necessary for the remainder of this lecture, where we only need the basic results.

**1.3.2 Notation:** For a multi-index  $\alpha = (\alpha_1, \dots, \alpha_d)$  with nonnegative integer  $\alpha_i$  and a function with sufficient differentiability, we define the derivative

$$\partial^\alpha f = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} f.$$

The order of  $\partial^\alpha$  is

$$|\alpha| = \sum \alpha_i.$$

**1.3.3 Definition:** If for a given function  $u$  there exists a function  $w$  such that

$$\int_{\Omega} w \varphi \, dx = - \int_{\Omega} u \partial_i \varphi \, dx, \quad \forall \varphi \in \mathcal{C}_{00}^\infty(\Omega), \quad (1.39)$$

then we define  $\partial_i u := w$  as the **distributional derivative** (partial) of  $u$  with respect to  $x_i$ . Here,  $\mathcal{C}_{00}^\infty(\Omega)$  is the space of all functions in  $C^\infty(\Omega)$  with compact support in  $\Omega$ .

Similarly through integration by parts, we define distributional directional derivatives, distributional gradients,  $\partial^\alpha u$ , etc.

We call a distributional derivative **weak derivative** in  $L^p$  if it is a function in this space.

**Remark 1.3.4.** Formula (1.39) is the usual integration by parts. Therefore, whenever  $u \in \mathcal{C}^1$  in a neighborhood of  $x$ , the distributional derivative and the usual derivative coincide.

**Example 1.3.5.** Let  $\Omega = \mathbb{R}$  and  $u(x) = |x|$ . Intuitively, it is clear that the distributional derivative, if it exists, must be the **Heaviside function**

$$w(x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0. \end{cases} \quad (1.40)$$

The proof that this is actually the distributional derivative is left to the reader.

**Example 1.3.6.** For the derivative of the Heaviside function in (1.40), we first observe that it must be zero whenever  $x \neq 0$ , since the function is continuously differentiable there. Now, we take a test function  $\varphi \in \mathcal{C}^\infty$  with support in the interval  $(-\varepsilon, \varepsilon)$  for some positive  $\varepsilon$ . Let  $w'(x)$  be the derivative of  $w$ . Then, by integration by parts

$$\int_{-\varepsilon}^{\varepsilon} w(x)\varphi'(x) dx = - \int_{-\varepsilon}^0 w(x)\varphi'(x) dx - \int_0^{\varepsilon} w(x)\varphi'(x) dx + 2\varphi(0) = 2\varphi(0),$$

since  $w'(x) = 0$  under both integrals. Thus,  $w'(x)$  is an object which is zero everywhere except at zero, but its integral against a test function  $\varphi$  is nonzero. This contradicts our notion, that integrable functions can be changed on a set of measure zero without changing the integral. Indeed,  $w'$  is not a function in the usual sense, and we write  $w'(x) = 2\delta(x)$ , where  $\delta(x)$  is the **Dirac  $\delta$ -distribution**, which is defined by the two conditions

$$\begin{aligned} \delta(x) &= 0, & \forall x \neq 0 \\ \int_{\mathbb{R}} \delta(x)\varphi(x) dx &= \varphi(0), & \forall \varphi \in \mathcal{C}^0(\mathbb{R}). \end{aligned}$$

We stress that  $\delta$  is not an integrable function, or a function at all.

**1.3.7 Definition:** The **Sobolev space**  $W^{k,p}(\Omega)$  is the space

$$W^{k,p}(\Omega) = \{u \in L^p(\Omega) \mid \partial^\alpha u \in L^p(\Omega) \forall |\alpha| \leq k\}, \quad (1.41)$$

where the derivatives are understood in weak sense. Its norm is defined by

$$\|v\|_{k,p}^p = \|v\|_{k,p;\Omega}^p = \sum_{|\alpha| \leq k} \|\partial^\alpha v\|_{L^p(\Omega)}^p. \quad (1.42)$$

The following seminorm will be useful:

$$|v|_{k,p}^p = |v|_{k,p;\Omega}^p = \sum_{|\alpha|=k} \|\partial^\alpha v\|_{L^p(\Omega)}^p. \quad (1.43)$$

**1.3.8 Notation:** We will use the notation

$$\|v\|_0 = \|v\|_{0;\Omega} = \|v\|_{L^2(\Omega)}.$$

Accordingly,  $W^{0,p}(\Omega) = L^p(\Omega)$ .

**1.3.9 Corollary:** There holds

$$W^{k,p}(\Omega) \subset W^{k-1,p}(\Omega) \subset \dots \subset W^{0,p}(\Omega) = L^p(\Omega) \quad (1.44)$$

**1.3.10 Definition:** The **Sobolev space**  $H^{k,p}(\Omega)$  is the completion of  $C^\infty(\Omega)$  with respect to the norm  $\|\cdot\|_{k,p}^p$ .  
In the case  $p = 2$ , we write  $H^k(\Omega) = H^{k,2}(\Omega)$ .

**1.3.11 Theorem (Meyers-Serrin):**

$$H^{k,p}(\Omega) \cong W^{k,p}(\Omega)$$

**Example 1.3.12.** Functions, which are in  $W^{k,p}(\Omega)$  or not.

1. The function  $x/|x|$  is in  $H^1(B_1(0))$  if  $d = 3$ , but not if  $d = 2$ .

**1.3.13 Definition:** A bounded domain  $\Omega \subset \mathbb{R}^d$  is said to have  $C^k$ -boundary or to be a  $C^k$ -domain, if there is a finite covering  $\{U_i\}$  of its boundary  $\partial\Omega$ , such that for each  $U_i$  there is a mapping  $\Phi_i \in C^k(U_i)$  with the following properties:

$$\begin{aligned} \Phi_i(\partial\Omega \cap U_i) &\subset \{\mathbf{x} \in \mathbb{R}^d \mid x_1 = 0\}, \\ \Phi_i(\Omega \cap U_i) &\subset \{\mathbf{x} \in \mathbb{R}^d \mid x_1 > 0\}. \end{aligned} \quad (1.45)$$

The domain is called Lipschitz, if such a construction exists with Lipschitz-continuous mappings.

**1.3.14 Definition:** We say that a normed vector space  $U \subset V$  is **continuously embedded** in another space  $V$ , in symbolic language

$$U \hookrightarrow V, \quad (1.46)$$

if the inclusion mapping  $U \ni x \mapsto x \in V$  is continuous, that is, there is a constant  $c$  such that

$$\|x\|_V \leq c\|x\|_U. \quad (1.47)$$

If the spaces  $U$  and  $V$  consist of equivalence classes, the inclusion may involve choosing representatives on the left or on the right.

**1.3.15 Theorem:** Let  $\Omega \subset \mathbb{R}^d$  be a bounded Lipschitz domain. For the space  $W^{k,p}(\Omega)$  define the number

$$s = k - \frac{d}{p}. \quad (1.48)$$

Assume  $k_1 \leq k_2$  and  $p_1, p_2 \in [1, \infty)$ . Then, if  $s_1 \geq s_2$ , we have the continuous embedding

$$W^{k_1, p_1}(\Omega) \hookrightarrow W^{k_2, p_2}(\Omega). \quad (1.49)$$

**1.3.16 Lemma:** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^d$ . Then, there exists a constant  $c$  only depending on  $\Omega$ , such that every function  $u \in H^1(\Omega) \cap C^1(\overline{\Omega})$  admits the estimate

$$\|u\|_{L^p(\partial\Omega)} \leq c\|u\|_{W^{1,p}(\Omega)}. \quad (1.50)$$

**1.3.17 Theorem (Trace theorem):** Let  $\Omega$  be a bounded Lipschitz domain in  $\mathbb{R}^d$ . Then, every function  $u \in W^{1,p}(\Omega)$  has a well defined trace  $\gamma u \in L^p(\partial\Omega)$  and there holds

$$\|\gamma u\|_{L^p(\partial\Omega)} \leq c\|u\|_{W^{1,p}(\Omega)}, \quad (1.51)$$

with the same constant as in the previous lemma. We simply write

$$u|_{\partial\Omega} = \gamma u. \quad (1.52)$$

**Remark 1.3.18.** The trace theorem guarantees that the imposition of Dirichlet boundary conditions on Sobolev functions is a reasonable operation. In particular, it ensures that  $H^1(\Omega)$  and  $H_0^1(\Omega)$  are indeed different spaces. The same does not hold, if we complete  $C^1(\Omega)$  and  $C_0^1(\Omega)$  in  $L^2(\Omega)$ .

Remarkably, we set out defining  $W^{1,p}(\Omega)$  as a subset of  $L^p(\Omega)$ , which consists of functions “defined up to a set of measure zero”. Now it turns out, that functions in  $W^{1,p}(\Omega)$  can have well-defined values on certain sets of measure zero. From the point of view of subsets of  $L^p(\Omega)$ , this is always to be understood by choosing representatives of the equivalence class. This is also the reason why we write “ $\hookrightarrow$ ” instead of “ $\subset$ ”.



**1.3.19 Definition:** A function  $f \in C^0(\Omega)$  is Hölder-continuous with exponent  $\gamma \in (0, 1]$ , if there is a constant  $C_f$  such that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq C_f |\mathbf{x} - \mathbf{y}|^\gamma \quad \forall \mathbf{x}, \mathbf{y} \in \Omega. \quad (1.53)$$

In particular, for  $\gamma = 1$ , we obtain Lipschitz-continuity.

We define the Hölder space  $C^{k,\gamma}$  of  $k$ -times continuously differentiable functions such that all derivatives of order  $k$  are Hölder-continuous. The norm is

$$\|u\|_{C^{k,\gamma}(\Omega)} = \max_{|\alpha| \leq k} \sup_{\mathbf{x}, \mathbf{y} \in \Omega} \frac{|\partial^\alpha u(\mathbf{x}) - \partial^\alpha u(\mathbf{y})|}{|\mathbf{x} - \mathbf{y}|^\gamma} \quad (1.54)$$

**1.3.20 Theorem:** Let  $\Omega \subset \mathbb{R}^d$  be a bounded Lipschitz domain. If  $s = k - \frac{d}{p} > j + \gamma$ , then every function in  $W^{k,p}$  has a representative in  $C^{j,\gamma}$ . We write

$$W^{k,p}(\Omega) \hookrightarrow C^{j,\gamma}(\Omega). \quad (1.55)$$

**1.3.21 Corollary:** Elements of Sobolev spaces are continuous if the derivative order is sufficiently high. In particular,

$$\begin{aligned} H^1(\Omega) &\hookrightarrow C(\Omega) & d = 1, \\ H^2(\Omega) &\hookrightarrow C(\Omega) & d = 2, 3. \end{aligned} \quad (1.56)$$

## 1.4 Regularity of Weak Solutions

**1.4.1.** So far, we have proven existence and uniqueness of weak solutions. We have seen, that these solutions may not even be continuous, far from differentiable. In this section, we collect a few results from the analysis of elliptic pde which establish higher regularity under stronger conditions.

**1.4.2 Definition:** The space  $W_{\text{loc}}^{k,p}(\Omega)$  consists of functions  $u$  such that  $u \in W^{k,p}(\Omega_1)$  for any  $\Omega_1 \subset\subset \Omega$ , where the latter reads compactly embedded, namely  $\overline{\Omega_1} \subset \Omega$ . Similarly, we define  $H_{\text{loc}}^k$ .

**1.4.3 Theorem ([Gilbarg and Trudinger, 1998, Theorem 8.8]):** Let  $a_{ij} \in C^{0,1}(\overline{\Omega})$  and  $b_i, c \in L^\infty(\Omega)$ . If  $u \in H^1(\Omega)$  is a solution to the elliptic equation and  $f \in L^2(\Omega)$ , then  $u \in H_{\text{loc}}^2(\Omega)$ .

**1.4.4 Theorem (Interior regularity):** Let  $a_{ij} \in C^{k,1}(\overline{\Omega})$  and  $b_i, c \in C^{k-1,1}(\overline{\Omega})$ . If  $u \in H^1(\Omega)$  is a solution to the elliptic equation and  $f \in W^{k,2}(\Omega)$ , then  $u \in H_{\text{loc}}^{k+2}(\Omega)$ .

*Proof.* [Gilbarg and Trudinger, 1998, Theorem 8.10] □

**1.4.5 Corollary:** If in the interior regularity theorem  $d = 2, 3$ , then  $u$  is a classical solution of the PDE if  $k \geq 2$ .  
If  $a_{ij}, b_i, c \in C^\infty(\overline{\Omega})$ , then  $u \in C^\infty(\Omega)$ .

**1.4.6 Theorem (Global regularity):** If in addition to the assumptions of the interior regularity theorem  $\Omega$  is a  $C^{k+2}$ -domain, then the solution  $u \in H_0^1(\Omega)$  to the homogeneous Dirichlet boundary value problem is in  $H^{k+2}(\Omega)$ .

*Proof.* [Gilbarg and Trudinger, 1998, Theorem 8.13] □

**1.4.7 Corollary:** Let  $\Omega \subset \mathbb{R}^d$  with  $d = 2, 3$  be a  $C^2$ -domain,  $a_{ij} \in C^{0,1}(\overline{\Omega})$  and  $b_i, c \in L^\infty(\Omega)$ . If  $u \in H_0^1(\Omega)$  is a solution to the elliptic equation and  $f \in L^2(\Omega)$ , then  $u \in H^2(\Omega)$ .

**1.4.8 Remark:** In order to guarantee a classical solution by these arguments, we must require that  $\Omega$  has  $C^4$  boundary.

**1.4.9 Remark:** The condition  $\partial\Omega \in C^2$  in the previous corollary can be replaced by the assumption that  $\Omega$  is convex.

**1.4.10 Theorem (Kondratev):** Let the assumptions of the interior regularity theorem Theorem 1.4.3 hold. Assume further that  $\partial\Omega$  is piecewise  $C^2$  with finitely many irregular points. Then, the solution  $u \in H_0^1(\Omega)$  of the elliptic PDE admits a representation

$$u = u_0 + \sum_{i=1}^n u_i, \quad (1.57)$$

where  $u_0 \in H^2(\Omega)$  and  $u_i$  is a singularity function associated with the irregular point  $\mathbf{x}_i$ .

*Proof.* [Kondrat'ev, 1967]

□

## Chapter 2

# Conforming Finite Element Methods

### 2.1 Meshes, shape functions, and degrees of freedom

**2.1.1 Definition:** Let  $T \subset \mathbb{R}^d$  be a polyhedron. We call the lower dimensional polyhedra constituting its boundary **facets**. A facet of dimension zero is called **vertex**, of dimension one **edge**, and a facet of codimension one is called a **face**.

**2.1.2 Definition:** A **mesh**  $\mathbb{T}$  is a nonoverlapping subdivision of the domain  $\Omega$  into polyhedral **cells** denoted by  $T$ , for instance simplices, quadrilaterals, or hexahedra. The faces of a cell are denoted by  $F$ , the vertices by  $\mathbf{X}$ . Cells are typically considered open sets. A mesh  $\mathbb{T}$  is called regular, if each face  $F \subset \partial T$  of the cell  $T \in \mathbb{T}$  is either a face of another cell  $T'$ , that is,  $\bar{F} = \bar{T} \cap \bar{T}'$ , or a subset of  $\partial\Omega$ .

**Remark 2.1.3.** For this introduction, we will assume that indeed  $\Omega$  is the union of mesh cells, which means, that its boundary consists of a finite union of planar faces. The more general case of a mesh approximating the domain will be deferred to later discussion.

**2.1.4 Definition:** With a mesh cell  $T$ , we associate a finite dimensional **shape function** space  $\mathcal{P}(T)$  of dimension  $n_T$ . The term **node functional** denotes linear functionals on this space.

A set of node functionals  $\{\mathcal{N}_T^i\}_{i=1,\dots,n_T}$  is called **unisolvent** on  $\mathcal{P}(T)$  if for any vector  $\mathbf{u} = (u_1, \dots, u_{n_T})^T$  there exists a unique  $u \in \mathcal{P}(T)$  such that

$$\mathcal{N}_T^i(u) = \mathbf{u}_i, \quad i = 1, \dots, n_T. \quad (2.1)$$

A **finite element** is a set of shape function spaces  $\mathcal{P}(T)$  for all  $T \in \mathbb{T}$  together with unisolvent set of node functionals.

**2.1.5 Notation:** If the node functionals  $\mathcal{N}^i$  are unisolvent on  $\mathcal{P}(T)$ , then, there is a basis  $\{p_k\}$  of  $\mathcal{P}(T)$  such that

$$\mathcal{N}^i(p_k) = \delta_{ik}. \quad (2.2)$$

We refer to  $\{p_k\}$  as **shape function basis** and use the term **degrees of freedom** for both the node functionals and the basis functions.

**2.1.6 Definition:** Node functionals can be associated with the cell  $T$  or with one of its lower dimensional boundary facets. We call this association the **topology** of the finite element.

**2.1.7 Definition:** The **finite element space** on the mesh  $\mathbb{T}$ , denoted by  $V_{\mathbb{T}}$  is a subset of the concatenation of all shape function spaces,

$$V_{\mathbb{T}} \subset \{f \in L^2(\Omega) \mid f|_T \in \mathcal{P}(T)\}. \quad (2.3)$$

The **degrees of freedom** of  $V_{\mathbb{T}}$  are the union of all node functionals, where we identify node functionals associated to boundary facets among all cells sharing this facet. The resulting dimension is

$$n = \dim V_{\mathbb{T}} \leq \sum n_T. \quad (2.4)$$

**2.1.8 Notation:** When we enumerate the degrees of freedom of  $V_{\mathbb{T}}$ , we obtain a global numbering of degrees of freedom  $\mathcal{N}^i$  with  $i = 1, \dots, n$ . For each mesh cell, we have a local numbering  $\mathcal{N}_T^j$  with  $j = 1, \dots, n_T$ . By construction of the finite element space, there is a unique  $i$ , such that  $\mathcal{N}_T^j(f) = \mathcal{N}^i(f)$  for all cells  $T$  and local indices  $j$ . The converse is not true due to the identification process.

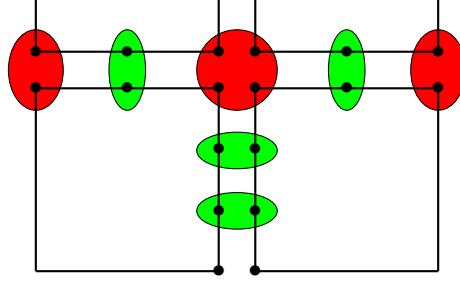


Figure 2.1: Identification of node functionals. The node functionals on shared edges (separated for presentation purposes) are distinguished locally as belonging to their respective cells, but identical global indices are assigned to all nodes in a single circle. Thus, all associated shape functions obtain the same coefficient in the global basis representation of a finite element function  $u$ .

**2.1.9 Definition:** We refer to the mapping between  $\mathcal{N}^i$  and  $\mathcal{N}_T^j$  as the mapping between global and local indices

$$\iota : (T, j) \mapsto i. \quad (2.5)$$

It induces a “natural” basis  $\{v_i\}$  of  $V_{\mathbb{T}}$  by

$$v_i|_T = p_{T,j}, \quad (2.6)$$

where  $\{p_{T,j}\}$  is the shape function basis on  $T$ . For each  $\mathcal{N}^i$ , we define  $\mathbb{T}(\mathcal{N}^i)$  as the set of cells  $T$  sharing the node functional  $\mathcal{N}^i$ , and

$$\Omega(\mathcal{N}^i) = \bigcup_{T \in \mathbb{T}(\mathcal{N}^i)} T. \quad (2.7)$$

**2.1.10 Lemma:** The support of the basis function  $v_i \in V_{\mathbb{T}}$  is

$$\text{supp}(v_i) \subset \Omega(\mathcal{N}^i).$$

**2.1.11 Lemma:** Let  $\mathbb{T}$  be a subdivision of  $\Omega$ , and let  $u$  be a function on  $\Omega$ , such that  $u|_T \in C^1(\bar{T})$  for each  $T \in \mathbb{T}$ . Then,

$$u \in H^1(\Omega) \iff u \in C(\bar{\Omega}). \quad (2.8)$$

**2.1.12 Lemma:** We have  $V_T \subset C(\bar{\Omega})$  if and only if for every facet  $F$  of dimension  $d_F < d$  there holds that

1. the traces of the spaces  $\mathcal{P}(T)$  on  $F$  coincide for all cells  $T$  having  $F$  as a facet,
2. The node functionals associated to the facet are unisolvent on this trace space.

### 2.1.1 Shape function spaces on simplices

**2.1.13 Definition:** A simplex  $T \in \mathbb{R}^d$  with vertices  $\mathbf{X}_0, \dots, \mathbf{X}_d$  is described by a set of  $d + 1$  **barycentric coordinates**  $\boldsymbol{\lambda} = (\lambda_0, \dots, \lambda_d)^T$  such that

$$0 \leq \lambda_i \leq 1 \quad i = 0, \dots, d; \quad (2.9)$$

$$\lambda_i(\mathbf{X}_j) = \delta_{ij} \quad i, j = 0, \dots, d \quad (2.10)$$

$$\sum \lambda_i(\mathbf{x}) = 1, \quad (2.11)$$

and there holds

$$T = \left\{ x \in \mathbb{R}^d \mid x = \sum \mathbf{X}_k \lambda_k \right\}. \quad (2.12)$$

**2.1.14 Lemma:** There is a matrix  $B_T \in \mathbb{R}^{d+1 \times d}$  and a vector  $b_T \in \mathbb{R}^{d+1}$ , such that

$$\boldsymbol{\lambda} = B_T \mathbf{x} + b_T. \quad (2.13)$$

**2.1.15 Corollary:** The barycentric coordinates  $\lambda_0, \dots, \lambda_d$  are the linear Lagrange interpolating functions for the points  $\mathbf{X}_0, \dots, \mathbf{X}_d$ . In particular,  $\lambda_k \equiv 0$  on the facet not containing  $\mathbf{X}_k$ .

**Example 2.1.16.** We can use barycentric coordinates to define interpolating polynomials on simplicial meshes easily, as in Table 2.1.

**Remark 2.1.17.** The functions  $\lambda_i(x)$  are the shape functions of the linear  $P_1$  element on  $T$ . They allow us to define basis functions on the cell  $T$  without use of a reference element  $\hat{T}$ .

Note that  $\lambda_i \equiv 0$  on the face opposite to the vertex  $x_i$ .

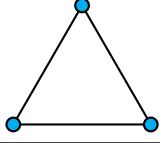
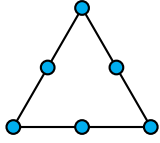
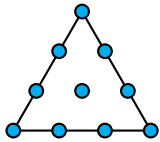
| Degrees of freedom  | Shape functions  |
|---|--|
|  | $\varphi_i = \lambda_i, \quad i = 0, 1, 2$   |
|  | $\varphi_{ii} = 2\lambda_i^2 - \lambda_i, \quad i = 0, 1, 2$<br>$\varphi_{ij} = 4\lambda_i\lambda_j \quad j \neq i$  |
|  | $\varphi_{iii} = \frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2) \quad i = 0, 1, 2$<br>$\varphi_{ij} = \frac{9}{2}\lambda_i\lambda_j(3\lambda_j - 1) \quad j \neq i$<br>$\varphi_0 = 27\lambda_0\lambda_1\lambda_2$ |

Table 2.1: Degrees of freedom and shape functions of simplicial elements in terms of barycentric coordinates

### 2.1.2 Shape functions on tensor product cells

**2.1.18 Definition:** The space of **tensor product polynomials** of degree  $k$  in  $d$  dimensions, denoted as  $\mathbb{Q}_k$  consists of polynomials of degree up to  $k$  in each variable. Given a basis for one-dimensional polynomials  $\{p_i\}_{i=0,\dots,k}$ , a natural basis for  $\mathbb{Q}_k$  is the **tensor product basis**

$$p_{i_1,\dots,i_d}(\mathbf{x}) = p_{i_1} \otimes \dots \otimes p_{i_d}(\mathbf{x}) = \prod_{k=1}^d p_{i_k}(x_k). \quad (2.14)$$

**Remark 2.1.19.** Note that the basis functions of  $\mathbb{Q}_k$  can be denoted as products of univariate polynomials, but that general polynomials in this space as linear combinations of these basis functions do not have this structure.



**2.1.20 Lemma:** Let  $\{\mathcal{N}_j\}$  be a set of one-dimensional node functionals dual to the one-dimensional basis  $\{p_i\}$  such that

$$\mathcal{N}_j(p_i) = \delta_{ij}. \quad (2.15)$$

Then, a dual basis for  $\{p_{i_1, \dots, i_d}\}$  is obtained by defining on the tensor product basis of  $\mathbb{Q}_k$

$$\mathcal{N}_{j_1, \dots, j_d}(p_{i_1, \dots, i_d}) = \mathcal{N}_{j_1} \otimes \dots \otimes \mathcal{N}_{j_d}(p_{i_1} \otimes \dots \otimes p_{i_d}) = \prod_{k=1}^d \mathcal{N}_{j_k}(p_{i_k}). \quad (2.16)$$

*Proof.* It is a theorem in linear algebra, that a linear functional on a vector space is uniquely defined by its values on a basis of the space. Thus, (2.16) uniquely defines the node functionals  $\mathcal{N}_{j_1, \dots, j_d}$ . The duality property follows from the fact that

$$\mathcal{N}_{j_1, \dots, j_d}(p_{i_1, \dots, i_d}) = \prod_{k=1}^d \delta_{i_k, j_k},$$

which is one if and only if all index pairs match and zero in all other cases.  $\square$

**Example 2.1.21.** Let a basis  $\{p_i\}$  of the univariate space  $\mathbb{P}_k$  be defined by Lagrange interpolation in  $k+1$  points  $t_j \in [0, 1]$ . A basis of the  $d$ -dimensional space  $\mathbb{Q}_k$  is then obtained by all possible products

$$p_{i_1, \dots, i_d}(\mathbf{x}) = \prod_{k=1}^d p_{i_k}(x_k).$$

The node functionals following the construction above are obtained by

$$\mathcal{N}_{j_1, \dots, j_d}(p_{i_1, \dots, i_d}) = \prod_{k=1}^d p_{i_k}(x_{j_k}).$$

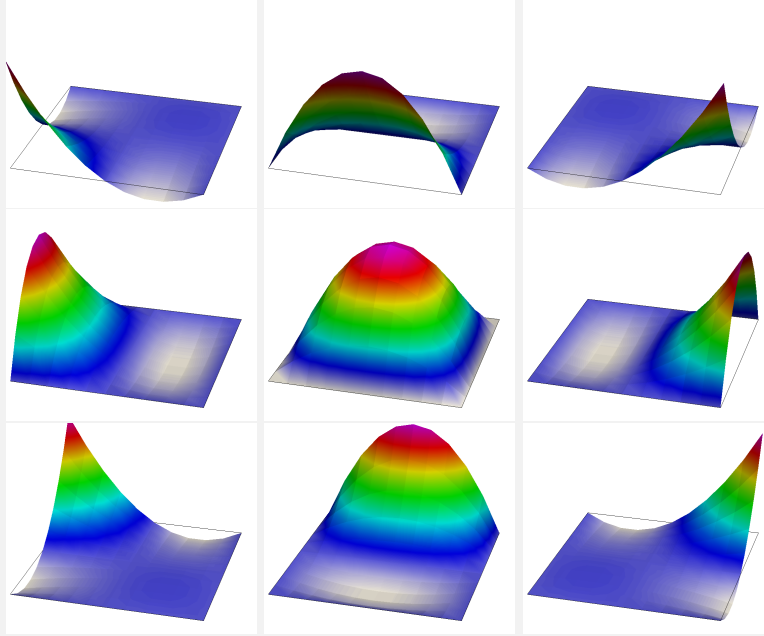
Finally, we have to convert the term on the right into an expression, which can be applied to any polynomial in  $\mathbb{Q}_k$ . To this end, we observe that

$$\prod_{k=1}^d p_{i_k}(x_{j_k}) = p_{i_1, \dots, i_d}(x_{j_1}, \dots, x_{j_d}).$$

Therefore, we conclude that the tensor product node functionals resulting from this construction are

$$\mathcal{N}_{j_1, \dots, j_d}(p) = p(x_{j_1}, \dots, x_{j_d}).$$

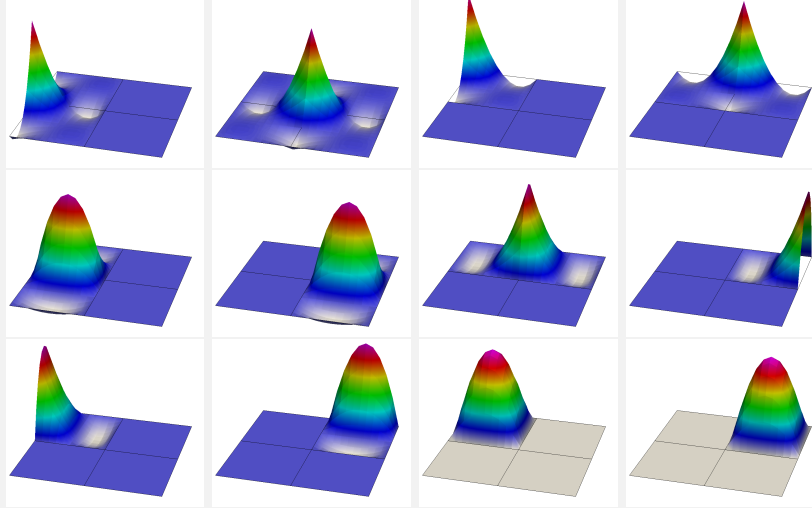
**2.1.22 Example (The space  $\mathbb{Q}_2$ ):**



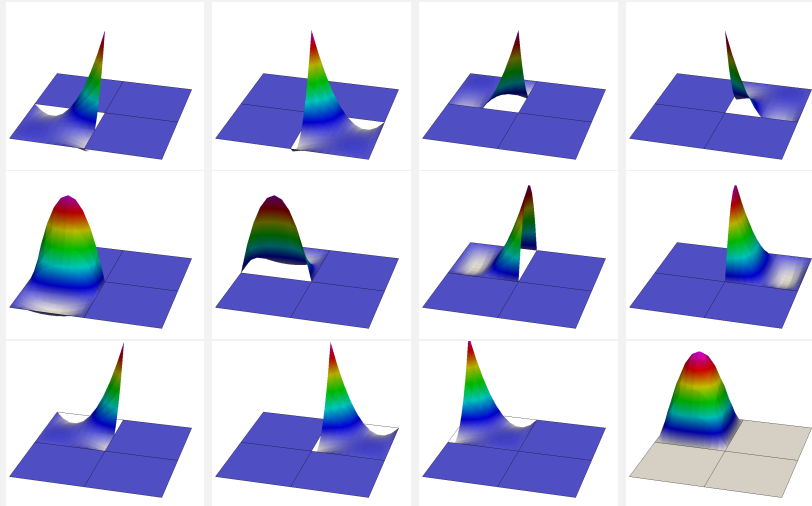
**2.1.23 Lemma:** The trace of the  $d$ -dimensional tensor product polynomial space  $\mathbb{Q}_k$  on the  $\delta$ -dimensional facets of the reference cube  $\hat{T} = (0, 1)^d$  is the  $\delta$ -dimensional space  $\mathbb{Q}_k$ .  
The traces from two cells sharing the same face coincide, if the mapping is continuous. Therefore, continuity can be achieved by unisolvent sets of node functionals on the face.

*Proof.* By keeping  $d - \delta$  variables constant in the tensor product basis in (2.14).  $\square$

**2.1.24 Example (Continuous basis functions):**



**2.1.25 Example (Discontinuous basis functions):**



**Example 2.1.26.** As a second example, we choose  $d = 2$  and the univariate space  $\mathbb{P}_2$  with node functionals

$$\mathcal{N}_0(p) = p(0), \quad \mathcal{N}_1(p) = \int_0^1 p(t) dt, \quad \mathcal{N}_2(p) = p(1), \quad (2.17)$$

that is, a mixture of Lagrange interpolation and orthogonality on the interval

$[0, 1]$ . The matching basis polynomials are

$$p_0(t) = 3(1-t)^2 - 2(1-t), \quad p_1(t) = 6t(1-t), \quad p_2(t) = 3t^2 - 2t. \quad (2.18)$$

Following the construction of the previous example, we obtain

$$\begin{aligned} \mathcal{N}_{00}(p) &= p(0, 0), & \mathcal{N}_{02}(p) &= p(0, 1), \\ \mathcal{N}_{20}(p) &= p(1, 0), & \mathcal{N}_{22}(p) &= p(1, 1). \end{aligned} \quad (2.19)$$

Then,

$$\mathcal{N}_{01}(p_{01}) = \mathcal{N}_{01}(p_0 \otimes p_1) = p_0(0) \int_0^1 p_1(y) dy = \int_0^1 p(0, y) dy. \quad (2.20)$$

Thus, the node functional  $\mathcal{N}_{01}$  is the integral over the left edge of the reference square. By the same construction,  $\mathcal{N}_{01}$  is the integral over the right edge.  $\mathcal{N}_{10}$  and  $\mathcal{N}_{12}$  are the integrals over the bottom and top edge, respectively. Finally,

$$\mathcal{N}_{11}(p_{11}) = \int_0^1 p_1(x) dx \int_0^1 p_1(y) dy = \int_0^1 \int_0^1 p_{11}(x, y) dx dy. \quad (2.21)$$

Thus, the tensor product of two line integrals becomes the integral over the area.

### 2.1.3 The Galerkin equations and Céa's lemma

**2.1.27 Definition (Galerkin approximation):** Let  $u \in V$  be determined by the weak formulation

$$a(u, v) = f(v) \quad \forall v \in V,$$

where  $V$  is a suitable function space including boundary conditions. The **Galerkin approximation**, also called **conforming approximation** of this problem reads as follows: choose a subspace  $V_n \subset V$  of dimension  $n$  and find  $u_n \in V_n$ , such that

$$a(u_n, v_n) = f(v_n) \quad \forall v_n \in V_n.$$

We will refer to this equation as the **discrete problem**.

**2.1.28 Corollary (Galerkin equations):** After choosing a basis  $\{v_i\}$  for  $V_n$ , the Galerkin equations are equivalent to a linear system

$$\mathbf{A} \mathbf{u} = \mathbf{f}, \quad (2.22)$$

with  $\mathbf{A} \in \mathbb{R}^{n \times n}$  and  $\mathbf{f} \in \mathbb{R}^n$  defined by

$$a_{ij} = a(v_j, v_i), \quad f_i = f(v_i). \quad (2.23)$$

**2.1.29 Lemma:** If the lemma of Lax-Milgram holds for  $a(.,.)$  on  $V$ , it holds on  $V_n \subset V$ . In particular, solvability of the Galerkin equations is implied.

**2.1.30 Lemma (C  a):** Let  $a(.,.)$  be a bounded and elliptic bilinear form on the Hilbert space  $V$ . Let  $u \in V$  and  $u_n \in V_n \subset V$  be the solution to the weak formulation and its Galerkin approximation

$$\begin{aligned} a(u, v) &= f(v) & \forall v \in V, \\ a(u_n, v_n) &= f(v_n) & \forall v_n \in V_n, \end{aligned}$$

respectively. Then, there holds

$$\|u - u_n\|_V \leq \frac{M}{\alpha} \inf_{v_n \in V_n} \|u - v_n\|_V. \quad (2.24)$$

**2.1.31 Lemma:** For a finite element discretization of Poisson's equation with the space  $V_{\mathbb{T}}$ , the Galerkin equations can be computed using the following formulas:

$$\begin{aligned} a_{ij} &= \int_{\Omega} \nabla v_j \cdot \nabla v_i \, dx = \int_{\Omega(\mathcal{N}^i)} \nabla v_j \cdot \nabla v_i \, dx = \sum_{T \in \mathbb{T}(\mathcal{N}^i)} \int_T \nabla v_j \cdot \nabla v_i \, dx \\ f_i &= \int_{\Omega} f v_i \, dx = \int_{\Omega(\mathcal{N}^i)} f v_i \, dx = \sum_{T \in \mathbb{T}(\mathcal{N}^i)} \int_T f v_i \, dx \end{aligned}$$

**2.1.32 Algorithm (Assembling the matrix):**

1. Start with a matrix  $\mathbf{A} = 0 \in \mathbb{R}^{n \times n}$
2. Loop over all cells  $T \in \mathbb{T}$
3. On each cell  $T$ , compute a cell matrix  $\mathbf{A}_T \in \mathbb{R}^{n_T \times n_T}$  by integrating

$$a_{T,ij} = \int_T \nabla p_{T,j} \cdot \nabla p_{T,i} \, dx, \quad (2.25)$$

where  $\{p_{T,i}\}$  is the shape function basis.

4. Assemble the cell matrices into the global matrix by

$$a_{\iota(i),\iota(j)} = a_{\iota(i),\iota(j)} + a_{T,ij} \quad i, j = 1, \dots, n_T. \quad (2.26)$$

### 2.1.4 Mapped finite elements

**2.1.33 Definition:** A mapped mesh  $\mathbb{T}$  is a set of cells  $T$ , which are defined by a single **reference cell**  $\hat{T}$  and individual smooth mappings

$$\begin{aligned}\Phi_T: \hat{T} &\rightarrow \mathbb{R}^d \\ \Phi_T(\hat{T}) &= T.\end{aligned}\tag{2.27}$$

The definition extends to small sets of reference cells, for instance for triangles and quadrilaterals.

**2.1.34 Example:** Let the reference triangle  $\hat{T}$  be defined by

$$\hat{T} = \left\{ \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} \middle| \hat{x}, \hat{y} > 0, \hat{x} + \hat{y} < 1 \right\}.\tag{2.28}$$

Then, every cell  $T$  spanned by the vertices  $\mathbf{X}_0$ ,  $\mathbf{X}_1$ , and  $\mathbf{X}_2$  is obtained by mapping  $\hat{T}$  by the affine mapping

$$\Phi_T(\hat{\mathbf{x}}) = \begin{pmatrix} X_1 - X_0 & X_2 - X_0 \\ Y_1 - Y_0 & Y_2 - Y_0 \end{pmatrix} \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} + \begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} =: \mathbf{B}_T \hat{\mathbf{x}} + \mathbf{b}_T\tag{2.29}$$

**2.1.35 Example:** The reference cell for a quadrilateral is the reference square  $\hat{T} = (0, 1)^2$ . Every quadrilateral  $T$  spanned by the vertices  $\mathbf{X}_0$  to  $\mathbf{X}_3$  is then obtained by the bilinear mapping

$$\Phi_T(\hat{\mathbf{x}}) = \mathbf{X}_0(1 - \hat{x})(1 - \hat{y}) + \mathbf{X}_1\hat{x}(1 - \hat{y}) + \mathbf{X}_2(1 - \hat{x})\hat{y} + \mathbf{X}_3\hat{x}\hat{y}\tag{2.30}$$

**2.1.36 Definition:** Mapped shape functions  $\{p_i\}$  on a mesh cell  $T$  are defined by a set of shape functions  $\{\hat{p}_i\}$  on the reference cell  $\hat{T}$  through **pull-back**

$$\begin{aligned}p_i(\mathbf{x}) &= \hat{p}_i(\Phi^{-1}(\mathbf{x})) = \hat{p}_i(\hat{\mathbf{x}}), \\ \nabla p_i(\mathbf{x}) &= \nabla \Phi^{-T}(\hat{\mathbf{x}}) \hat{\nabla} \hat{p}_i(\hat{\mathbf{x}})\end{aligned}\tag{2.31}$$

**2.1.37 Lemma:** Let  $\hat{T}$  be the reference triangle and let  $T$  be a triangular mesh cell with mapping  $\mathbf{x} = \Phi_T(\hat{\mathbf{x}}) = \mathbf{B}\hat{\mathbf{x}} + \mathbf{b}$ . Let there hold  $u(\mathbf{x}) = \hat{u}(\hat{\mathbf{x}})$ . Then,  $u \in H^k(T)$  if and only if  $\hat{u} \in H^k(\hat{T})$  and we have with some constant  $c$  the estimates

$$\begin{aligned} |\hat{u}|_{k,\hat{T}} &\leq c \|\mathbf{B}\|^k (\det \mathbf{B})^{-1/2} |u|_{k,T}, \\ |u|_{k,T} &\leq c \|\mathbf{B}^{-1}\|^k (\det \mathbf{B})^{1/2} |\hat{u}|_{k,\hat{T}}. \end{aligned} \quad (2.32)$$

**2.1.38 Lemma:** For a cell  $T$ , let  $R$  be the radius of the circumscribed circle and  $\varrho$  the radius of the inscribed circle. Then,

$$\|\mathbf{B}\| \leq cR, \quad \|\mathbf{B}^{-1}\| \leq c\varrho^{-1}. \quad (2.33)$$

**2.1.39 Assumption:** For more general mappings  $\Phi: \hat{T} \rightarrow T$ , we make the assumption, that they can be decomposed into three factors,

$$\Phi = \Phi_O \circ \Phi_S \circ \Phi_W, \quad (2.34)$$

where  $\Phi_O$  is a combination of translation and rotation,  $\Phi_S$  is a scaling with a characteristic length  $h_T$ , and  $\Phi_W$  is a warping function not changing the characteristic length.

**Example 2.1.40.** We construct the inverse of  $\Phi$  in two dimensions by the following three steps, using as  $h_T$  the length of the longest edge of  $T$ .

1. Choose  $\Phi_O$  as the rigid body movement which maps the longest edge to the interval  $(0, h_T)$  on the  $x$ -axis and the cell itself to  $T_O$  in the positive half plane. This mapping has the structure

$$\Phi_O^{-1}(\mathbf{x}) = \mathbf{S}\mathbf{x} - \mathbf{S}\mathbf{X}_0,$$

where  $\mathbf{S}$  is an orthogonal matrix and  $\mathbf{X}_0$  is the vertex moved to the origin.

2. Choose the scaling

$$\Phi_S^{-1}(\mathbf{x}) = \frac{1}{h_T} \mathbf{x},$$

such that the longest edge of the resulting cell  $T_S$  has the longest edge equal to the interval  $(0, 1)$  on the  $x$ -axis.

3. Warp the cell  $T_S$  into the reference cell  $\hat{T}$  by the mapping  $\Phi_W^{-1}$ . This operation leaves the longest edge untouched. For triangles, it is the uniquely defined linear transformation mapping the vertex not on the longest edge to  $(0, 1)$ . For quadrilaterals, it is a bilinear transformation.

In the first step, we have assumed that the cell is convex, which is always true for triangles. For nonconvex quadrilaterals, it can be shown that the determinant of  $\nabla\Phi$  changes sign inside the cell, such that these cells are not useful for computations.

The idea of this decomposition is, that we separate mappings changing the position, size, and shape of the cells.

**2.1.41 Lemma (Scaling lemma):** Let the typical length of a cell  $T$  be  $h_T$ . Assume there are constants  $0 < M_T, m_T, d_T, D_T$ , such that

$$\begin{aligned} \|\nabla\Phi_W(\hat{\mathbf{x}})\| &\leq M_T, \\ \|\nabla\Phi_W^{-1}(\hat{\mathbf{x}})\| &\leq m_T^{-1}, \\ d_T^2 &\leq \det \nabla\Phi_W(\hat{x}) \leq D_T^2. \end{aligned} \tag{2.35}$$

for all  $\hat{\mathbf{x}} \in \hat{T}$ . Then, for  $k = 0, 1$  and a constant  $c$

$$\begin{aligned} |\hat{u}|_{k,\hat{T}} &\leq c \frac{M_T}{d_T} h_T^{k-d/2} |u|_{k,T}, \\ |u|_{k,T} &\leq c \frac{D_T}{m_T} h_T^{d/2-k} |\hat{u}|_{k,\hat{T}}. \end{aligned} \tag{2.36}$$

This extends to higher derivatives under assumptions on higher derivatives of  $\Phi_T$ .

*Proof.* By the chain rule,  $\nabla\Phi_T = \nabla\Phi_O \nabla\Phi_S \nabla\Phi_W$ . By construction,  $\nabla\Phi_O$  is an orthogonal matrix, such that

$$\|\nabla\Phi_O\| = \|\nabla\Phi_O^{-1}\| = 1.$$

Since it preserves angles and lengths,  $\det \nabla\Phi_O = 1$ . Since  $\Phi_S$  is a multiple of the identity, we have

$$\|\nabla\Phi_S\| = h_T, \quad \|\nabla\Phi_S^{-1}\| = \frac{1}{h_T}, \quad \det \nabla\Phi_S = h_T^d.$$

By change of variables, we have

$$\int_T u^2 \, d\mathbf{x} = \int_{\hat{T}} \hat{u}^2 |\det \nabla\Phi_T| \, d\hat{\mathbf{x}} = \int_{\hat{T}} \hat{u}^2 \det \nabla\Phi_S \det \nabla\Phi_O \det \nabla\Phi_W \, d\hat{\mathbf{x}},$$

such that the case  $k = 0$  is proven immediately by

$$h_T^d d_T^2 \int_{\hat{T}} \hat{u}^2 \, d\hat{\mathbf{x}} \leq \int_T u^2 \, d\mathbf{x} \leq h_T^d D_T^2 \int_{\hat{T}} \hat{u}^2 \, d\hat{\mathbf{x}}$$



By the chain rule, we have

$$\widehat{\nabla} \widehat{u}(\widehat{\mathbf{x}}) = \nabla \Phi^T \nabla u(\mathbf{x}) = \nabla \Phi_W^T \nabla \Phi_S^T \nabla \Phi_O^T \nabla u(\mathbf{x}),$$

such that there holds

$$\begin{aligned} |\widehat{\nabla} \widehat{u}(\widehat{\mathbf{x}})| &\leq \|\nabla \Phi_W\|_{h_T} |\nabla u|, \\ |\nabla u(\mathbf{x})| &\leq \|\nabla \Phi_W^{-1}\|_{h_T^{-1}} |\widehat{\nabla} \widehat{u}|. \end{aligned} \tag{2.37}$$

□

**2.1.42 Remark:** We have  $d_T = D_T$ , if and only if the mapping is affine. The quotient  $M_T/m_T$  measures how much the shape of the mesh cell deviates from the reference cell. For instance, it is one for squares.

## 2.2 A priori error analysis

**2.2.1.** In this section, we develop error estimates of the following type

If the size of mesh cells converges to zero, then the difference between the true solution and the finite element solution converges to zero with a certain order.

They are thus a prediction, that the solutions actually converge, and they measure the asymptotic convergence rate. They do contain unknown constants, such that they are no prediction of the error on a given mesh.

The theory in this chapter is about bilinear forms which are bounded and elliptic on a subspace  $V \subset H^1(\Omega)$  determined by boundary conditions. We will choose  $V = H_0^1(\Omega)$ , even if more general boundary conditions are possible. It is a good approach to think of the Dirichlet problem for the Laplacian, even if we allow for somewhat more general equations.

### 2.2.1 Approximation of Sobolev spaces by finite elements

**2.2.2 Lemma (Poincaré inequality):** Let  $\Omega$  be a bounded Lipschitz domain. For any function  $u \in H^1(\Omega)$  define

$$\bar{u} = \frac{1}{|\Omega|} \int_{\Omega} u(\mathbf{x}) \, d\mathbf{x}, \quad (2.38)$$

where  $|\Omega|$  denotes the measure of  $\Omega$ . There exists a constant  $c$  depending on the domain only, such that each of the following inequalities hold:

$$\|u - \bar{u}\|_{L^2(\Omega)} \leq c \|\nabla u\|_{L^2(\Omega)} \quad (2.39)$$

$$\|u\|_{L^2(\Omega)}^2 \leq c \left( \|\nabla u\|_{L^2(\Omega)}^2 + \bar{u}^2 \right) \quad (2.40)$$

*Proof.* The proof exceeds the tools we have developed in this class. The proof in [Gilbarg and Trudinger, 1998, Section 7.8] seems elementary and direct, but is technical and requires star-shaped domains. The proof in [Evans, 1998, Section 5.8.1] is more elegant, but it uses compact embedding and is indirect, such that the constant cannot be determined.  $\square$

**2.2.3 Lemma (Bramble-Hilbert):** Let  $T \subset \mathbb{R}^d$  be a domain with Lipschitz boundary and let  $s(\cdot)$  be a bounded sublinear functional on  $H^{k+1}(T)$  with the property

$$s(p) = 0 \quad \forall p \in \mathbb{P}_k. \quad (2.41)$$

Then, there exists a constant  $c$  only dependent on  $T$  such that

$$|s(v)| \leq c |v|_{k+1,T}. \quad (2.42)$$

*Proof.* Since  $s(\cdot)$  is sublinear and vanishes on  $\mathbb{P}_k$ , we have for  $v \in H^{k+1}(T)$ :

$$|s(v)| \leq |s(v+p)| + |s(p)| = |s(v+p)| \quad \forall p \in \mathbb{P}_k. \quad (2.43)$$

We will construct a polynomial, such that

$$\overline{\partial^\alpha(v+p)} = \frac{1}{|T|} \int_T \partial^\alpha(v+p) \, dx = 0 \quad \forall |\alpha| \leq k, \quad (2.44)$$

that is, the sum  $v+p$  and all its derivatives up to order  $k$  are mean-value free. Thus, by recursive application of Poincaré inequality, we get for  $|\alpha| \leq k$

$$\begin{aligned} \|v+p\|_{L^2(T)}^2 &\leq c \left[ \|\nabla(v+p)\|_{L^2(T)}^2 + \overline{v+p}^2 \right] \leq c |v+p|_{1,T}^2 \\ \|\partial^\alpha(v+p)\|_{L^2(T)}^2 &\leq c \left[ \|\nabla \partial^\alpha(v+p)\|_{L^2(T)}^2 + \overline{\partial^\alpha(v+p)}^2 \right] \leq c |v+p|_{|\alpha|+1,T}^2, \end{aligned}$$

such that  $\|v + p\|_{k+1;T} \leq |v + p|_{k+1;T}$ . Furthermore, since  $p \in \mathbb{P}_k$

$$\|\partial^\alpha(v + p)\|_{L^2(T)} = \|\partial^\alpha v\|_{L^2(T)} \quad \forall |\alpha| = k + 1.$$

Combining with (2.43), we obtain

$$|s(v)| \leq c|v + p|_{k+1;T} \leq c|v|_{k+1;T}.$$

It remains to construct the polynomial  $p$  with the desired properties. To this end, we note that for two multi-indices  $\alpha$  and  $\beta$  holds that  $\partial^\alpha \mathbf{x}^\beta = 0$  as soon as  $\alpha_i > \beta_i$  for some index  $i$ . Let

$$p(\mathbf{x}) = \sum_{|\beta| \leq k} a_\beta \mathbf{x}^\beta.$$

Then, for any  $|\alpha| = k$  we get

$$\partial^\alpha p(\mathbf{x}) = \alpha! \delta_{\alpha\beta},$$

where  $\alpha! = \alpha_1! \alpha_2! \dots \alpha_d!$ . Thus, we can use condition (2.44) to fix the coefficients  $a_\beta$  to

$$a_\beta = \frac{1}{\beta! |T|} \int_T \partial^\beta v \, dx \quad |\beta| = k.$$

Thus, we have decomposed  $p = \tilde{p}_k + p_{k-1}$ , where  $\tilde{p}_k$  is known and  $p_{k-1} \in \mathbb{P}_{k-1}$ . Thus, we can repeat the process determining the coefficients of highest order in  $p_{k-1}$ ,

$$a_\beta = \frac{1}{\beta! |T|} \int_T \partial^\beta (v - \tilde{p}_k) \, dx \quad |\beta| = k - 1.$$

Recursion down to  $k = 0$  yields the polynomial  $p$  with the desired property.  $\square$

**2.2.4 Corollary:** Let  $\Pi: H^{k+1}(T) \rightarrow \mathbb{P}_k$  be a continuous, linear projector. For any  $m \leq k$  there exists a constant  $c$  such that

$$\|u - \Pi u\|_{m,T} \leq c|u|_{k+1,T}. \quad (2.45)$$

**2.2.5 Definition:** Let  $\{\mathbb{T}_h\}$  for  $h > 0$  be a family of meshes parametrized by the parameter

$$h = \max_{T \in \mathbb{T}_h} h_T, \quad (2.46)$$

where  $h_T$  is the characteristic length from the discussion of mappings, for instance the diameter of  $T$ . Such a family is called **shape regular**, if the constants  $M_T$ ,  $m_T$ ,  $d_T$ , and  $D_T$  in the scaling lemma can be chosen independent of the cell  $T \in \mathbb{T}_h$  and of  $h > 0$ . The family is called **quasi-uniform**, if in addition there is a positive constant independent of  $h$  such that

$$h \leq c \min_{T \in \mathbb{T}_h} h_T. \quad (2.47)$$

**2.2.6 Definition:** Let  $V$  be a function space on  $\Omega$ , and let  $V_h = V_{\mathbb{T}_h}$  be a finite element space on the mesh  $\mathbb{T}_h$  on  $\Omega$  with node functionals  $\mathcal{N}_i$  and dual basis  $p_i$ , where  $i = 1, \dots, n_h$ . We define the **nodal interpolation** operator by

$$\begin{aligned} I_h: V &\rightarrow V_h \\ v &\mapsto \sum \mathcal{N}_i(v) p_i. \end{aligned} \quad (2.48)$$

**2.2.7 Lemma:** The nodal interpolation operator  $I_h$  is a projector. It is continuous on  $H^2(\Omega)$  if the dimension is  $d = 2, 3$  and the node functionals are defined as Lagrange interpolation.

**2.2.8 Definition:** On a mesh  $\mathbb{T}_h$ , we define the **broken Sobolev norm** and seminorm by

$$\begin{aligned} \|u\|_{k;h}^2 &= \sum_{T \in \mathbb{T}_h} \|u\|_{k;T}^2 \\ |u|_{k;h}^2 &= \sum_{T \in \mathbb{T}_h} |u|_{k;T}^2 \end{aligned} \quad (2.49)$$

**2.2.9 Theorem:** Let  $\{\mathbb{T}_h\}$  be a shape regular family of meshes. Let the finite element spaces  $V_h = V_{\mathbb{T}_h}$ . Let the nodal interpolation operator  $I_h$  be surjective onto  $\mathbb{P}_k$  on every cell  $T \in \mathbb{T}_h$  and continuous on  $H^{k+1}(\Omega)$ . Then, there is a constant  $c$  such that for any  $u \in H^{k+1}(\Omega)$  and  $m \leq k+1$  there holds

$$\|u - I_h u\|_{m;h} \leq ch^{k+1-m} |u|_{k+1;h}. \quad (2.50)$$

*Proof.* We have by definition

$$|u - I_h u|_{m;h} = \sum_{T \in \mathbb{T}_h} |u - I_h u|_{m;T}.$$

Using the scaling lemma, we get

$$|u - I_h u|_{m;T} \leq ch_T^{d/2-m} |\widehat{u} - \widehat{I_h u}|.$$

On the reference cell, we use the Bramble-Hilbert lemma, more precisely, Corollary 2.2.4 to obtain

$$|\widehat{u} - \widehat{I_h u}| \leq c |\widehat{u}|_{k+1;\widehat{T}}.$$

Scaling back yields

$$|\widehat{u}|_{k+1;\widehat{T}} \leq ch_T^{k+1-d/2} |u|_{k+1;T}.$$

Combining, we obtain

$$|u - I_h u|_{m;T} \leq ch_T^{k+1-d/2-m+d/2} |u|_{k+1;T}.$$

Summing up and pulling the maximum of  $h_T^{k+1-m}$  out of the sum yields the result for  $h_T \leq 1$ .  $\square$

**Remark 2.2.10.** Strictly speaking, we have only proven the result for  $h_T \leq 1$ . But then, if  $\text{diam } \Omega = 1$ , this condition is always true. Therefore, by rescaling the domain before computing, the estimate holds in general.

**2.2.11 Corollary:** Let  $a(.,.)$  be a bounded and elliptic bilinear form on  $V = H_0^1(\Omega)$  and let the finite element space  $V_h$  be defined on a shape-regular family of meshes  $\{\mathbb{T}_h\}$ , such that the interpolation estimate eq. (2.50) holds. If furthermore the solution  $u \in H^{k+1}(\Omega)$ , the error of the finite element solution  $u_h \in V_h \subset V$  admits the estimate

$$\|u - u_h\|_{1;h} \leq ch^k |u|_{k+1;h}. \quad (2.51)$$

**2.2.12.** For 2nd order elliptic problems, we have now derived estimates of the  $H^1$ -norm of the error under the assumption that the solution exhibits further regularity. Now, let us drop this assumption for an assumption on the boundary condition only, to obtain a qualitative convergence result.

**2.2.13 Theorem:** Let  $a(.,.)$  be a bounded and elliptic bilinear form on  $V = H_0^1(\Omega)$  and let the finite element space  $V_h$  be defined on a shape-regular family of meshes  $\{\mathbb{T}_h\}$ , such that the interpolation estimate eq. (2.50) holds. Let  $u \in V$  and  $u_h \in V_h$  be solutions to the exact and the finite element versions of a 2nd order boundary value problem. Then,

$$\lim_{h \searrow 0} \|u - u_h\|_{1;\Omega} = 0. \quad (2.52)$$

## 2.2.2 Estimates of stronger norms

**2.2.14.** So far, we have seen error estimates in a “natural norm” defined as a norm such that the Lax-Milgram lemma holds for a given bilinear form  $a(.,.)$ . In this subsection, we now consider the question of estimates in stronger norms, such that the bilinear form is not elliptic with respect to this norm.

**2.2.15 Definition:** Let  $\|\cdot\|_X$  and  $\|\cdot\|_Y$  be norms on a vector space  $V$ . We call  $\|\cdot\|_X$  a **stronger norm** than  $\|\cdot\|_Y$ , if there is a constant  $c$  such that for all  $v \in V$ :

$$\|v\|_Y \leq c\|v\|_X. \quad (2.53)$$

In this case,  $\|\cdot\|_Y$  is called the **weaker norm**. If a converse inequality holds, the norms are called **equivalent**.

**Example 2.2.16.** For a bounded domain  $\Omega$ , the norms  $\|\cdot\|_{k+1}$  and  $\|\cdot\|_k$  are both defined on  $V = H_0^1(\Omega)$ . By the Sobolev embedding theorem, there is a constant  $c$  such that for any  $v \in V$

$$\|v\|_k \leq c\|v\|_{k+1}$$

**2.2.17 Lemma (Inverse estimate):** Let  $T$  be a mesh cell of size  $h_T$ . Then, there is a constant only depending on  $k$  and the constants of the scaling lemma, such that for every  $u \in \mathbb{P}_k$  there holds

$$\|u\|_{1;T} \leq ch_T^{-1} \|u\|_{0;T}. \quad (2.54)$$

**2.2.18 Theorem:** Let  $a(.,.)$  be a bounded and elliptic bilinear form on  $V \subset H^1(\Omega)$  and let  $\{\mathbb{T}_h\}$  be a family of quasi-uniform meshes with finite element spaces  $V_h \subset V$  containing the space  $\mathbb{P}_k$  with  $k \geq 2$  on each mesh cell. If furthermore the solution  $u \in H^{k+1}(\Omega)$ , the error between the exact and the finite element solution to a uniquely solvable elliptic boundary value problem, admits the estimate

$$\|u - u_h\|_{2;h} \leq ch^{k-1}|u|_{k+1;h}. \quad (2.55)$$

*Proof.* We cannot apply Céa's lemma directly, since the bilinear form is not elliptic with respect to the broken  $H^2$ -inner product on the space  $H^1(\Omega)$ . On the other hand, the error is not polynomial, such that we cannot apply the inverse estimate to it. Instead, we use triangle inequality

$$\|u - u_h\|_{2;h} \leq \|u - I_h u\|_{2;h} + \|I_h u - u_h\|_{2;h},$$

and observe, that we already have the desired estimate for the first term. For the second, we estimate by inverse estimate on each cell

$$\min_{T \in \mathbb{T}_h} h_T \|I_h u - u_h\|_{2;h} \|I_h u - u_h\|_1 \leq c \|I_h u - u_h\|_1^2,$$

and

$$\begin{aligned} \|I_h u - u_h\|_1^2 &\leq \frac{c}{\gamma} a(I_h u - u_h, I_h u - u_h) \\ &= \frac{c}{\gamma} a(I_h u - u, I_h u - u_h) \\ &\leq \frac{cM}{\gamma} \|I_h u - u\|_1 \|I_h u - u_h\|_1 \\ &\leq Ch^k |u|_{k+1;h} \|I_h u - u_h\|_1, \end{aligned}$$

where we have used Galerkin orthogonality and the interpolation estimate. Combining the two and using the quasi-uniformity, we obtain the result of the theorem.  $\square$

### 2.2.3 Estimates of weaker norms and linear functionals

**2.2.19.** Deriving optimal error estimates in the  $L^2$ -norm cannot be achieved by the same technique as used for the  $H^2$ -norm, as the following simple argument shows: using triangle inequality, we obtain

$$\|u - u_h\|_{L^2} \leq \|u - I_h u\|_{L^2} + \|I_h u - u_h\|_{L^2},$$

we obtain that the first term is of order  $h^{k+1}$ . Thus, for an optimal error estimate, we require that the second term be of order  $h^{k+1}$  as well. We need

a replacement of the inverse estimate, which gains an order of  $h$  instead of losing it. This is Poincaré inequality, but it requires an almost mean-value free function. And since  $I_h u$  is not the interpolant of  $u_h$ , we cannot guarantee a small mean value of the difference on each mesh cell.

To the rescue comes a “duality argument” known as Aubin-Nitsche trick, which we introduce now.

**2.2.20 Definition:** Let the weak form of a boundary value problem on the domain  $\Omega$  be defined as: find  $u \in V$  such that

$$a(u, v) = f(v) \quad \forall v \in V.$$

Then, the **dual problem**, also called **adjoint problem**, with right hand side  $g \in V^*$  is: find  $u^* \in V$  such that

$$a(v, u^*) = g(v) \quad \forall v \in V.$$

**2.2.21 Lemma:** The adjoint problem of the Dirichlet boundary value problem for Poisson’s equation is equal to the dual problem, that is, for  $u \in H_0^1(\Omega)$ , the two statements

$$\begin{aligned} a(u, v) &= f(v) & \forall v \in V, \\ a(v, u) &= f(v) & \forall v \in V, \end{aligned}$$

are equivalent.

**2.2.22 Assumption (Elliptic regularity):** Let  $a(., .)$  be a bounded and elliptic bilinear form on  $H_0^1(\Omega)$ . We say that a boundary value problem has **elliptic regularity**, if for any  $g \in L^2(\Omega)$  the solution  $u$  is in  $H^2(\Omega)$ . In other words, there is a constant  $c$  independent of  $g$ , such that

$$\|u\|_{2;\Omega} \leq c\|g\|_{0;\Omega}. \quad (2.56)$$

**Example 2.2.23.** By Remark 1.4.9, a second order PDE with coefficients  $a_{ij} \in C^{0,1}(\overline{\Omega})$  and  $b_i, c \in L^\infty(\Omega)$  has elliptic regularity. The same holds by integration by parts for its adjoint.

The same boundary value problem does not have elliptic regularity, if the domain has a nonconvex corner, since we have corner singularity functions which are not in  $H^2(\Omega)$ , and we can construct a right hand side  $g \in L^2(\Omega)$ , which produces such singularities.



**2.2.24 Theorem:** Let  $a(.,.)$  be a bounded and elliptic bilinear form on  $V = H_0^1(\Omega)$  and let the finite element space  $V_h$  be defined on a shape-regular family of meshes  $\{\mathbb{T}_h\}$ , such that the interpolation estimate (2.50) holds. If furthermore the solution  $u \in H^{k+1}(\Omega)$  and the dual problem has elliptic regularity, the error of the finite element solution  $u_h \in V_h \subset V$  admits the estimate

$$\|u - u_h\|_0 \leq ch^{k+1}|u|_{k+1;h}. \quad (2.57)$$

**2.2.25 Corollary:** Under the assumptions of Theorem 2.2.24, let  $J(.)$  be a bounded linear functional on  $L^2(\Omega)$ . Then,

$$|J(u) - J(u_h)|_0 \leq ch^{k+1}|u|_{k+1;h}. \quad (2.58)$$

## 2.2.4 Green's function and maximum norm estimates

**2.2.26.** In this section, we will introduce the basic concepts needed for pointwise error estimation. They heavily rely on Green's function, which is useful for a general understanding of the solution structure as well.

**2.2.27 Definition:** For a differential equation  $Lu = f$  in  $\Omega$  with boundary conditions  $u = 0$  on the boundary  $\partial\Omega$ , we define **Green's function**  $G(\mathbf{y}, \mathbf{x})$  associated to the point  $\mathbf{y} \in \Omega$  as solution to the problem

$$\begin{aligned} LG(\mathbf{y}, \mathbf{x}) &= \delta(\mathbf{x} - \mathbf{y}) & \forall \mathbf{x} \in \Omega, \\ G(\mathbf{y}, \mathbf{x}) &= 0 & \forall \mathbf{x} \in \partial\Omega. \end{aligned} \quad (2.59)$$

**2.2.28 Theorem:** Green's function for Poisson's equation on the whole space  $\Omega = \mathbb{R}^d$  is

$$G(\mathbf{y}, \mathbf{x}) = \begin{cases} -\frac{1}{2\pi} \log|\mathbf{x} - \mathbf{y}| & d = 2 \\ \frac{1}{d(d-2)|B_1(0)|} \frac{1}{|\mathbf{x} - \mathbf{y}|^{d-2}} & d \geq 3. \end{cases} \quad (2.60)$$

*Proof.* See [Evans, 1998, Section 2.2]. □

**2.2.29 Lemma:** Let  $\Omega \subset \mathbb{R}^d$  be a bounded domain. Then, Green's function associated to a point  $\mathbf{y} \in \Omega$  for a linear differential operator  $L$  on  $\Omega$  is obtained as the sum

$$G(\mathbf{y}, \mathbf{x}) = G_\infty(\mathbf{y}, \mathbf{x}) - G_0(\mathbf{y}, \mathbf{x}), \quad (2.61)$$

where  $G_\infty(\mathbf{y}, \mathbf{x})$  is Green's function for the whole space and  $G_0(\mathbf{y}, \mathbf{x})$  solves  $LG_0(\mathbf{y}, \mathbf{x}) = 0$  with boundary values  $G_\infty(\mathbf{y}, \mathbf{x})$ . If the domain is convex, then,  $G_0(\mathbf{y}, \cdot) \in H^2(\Omega)$  for any interior point  $\mathbf{y}$ .

**2.2.30 Theorem:** Let  $f \in C(\overline{\Omega})$  and let  $\Omega$  be such that the solution to

$$\begin{aligned} Lu &= f & \text{in } \Omega, \\ u &= 0 & \text{on } \partial\Omega, \end{aligned} \quad (2.62)$$

is in  $C^2(\overline{\Omega})$ . Then,

$$u(\mathbf{x}) = \int_{\Omega} f(\mathbf{y})G(\mathbf{y}, \mathbf{x}) \, d\mathbf{y}. \quad (2.63)$$

*Proof.* See [Evans, 1998, Section 2.2] □

**2.2.31 Theorem:** Let  $a(\cdot, \cdot)$  be a bounded and elliptic bilinear form on  $V = H_0^1(\Omega)$  on a bounded, convex domain  $\Omega \subset \mathbb{R}^2$ . Let the finite element space  $V_h$  be defined on a quasi-uniform family of meshes  $\{\mathbb{T}_h\}$  with local spaces  $\mathbb{P}_1$  or  $\mathbb{Q}_1$ . If furthermore the solution  $u \in W^{2,\infty}(\Omega)$ , the error of the finite element solution  $u_h \in V_h \subset V$  admits the estimate

$$\|u - u_h\|_\infty \leq ch^2(1 + |\log h|)|u|_{2,\infty}. \quad (2.64)$$

*Proof.* The proof relies on solving a dual problem for the error in a single point. The solution is Green's function for the adjoint equation, which is very irregular at the point of interest. Then, complicated analysis is needed to generate useful approximation estimates in spite of the singularity. Details can be found in [Schatz and Wahlbin, 1977, Rannacher and Scott, 1982]. □

**Remark 2.2.32.** This estimate extends to  $\Omega \in \mathbb{R}^d$  for  $d \geq 3$  and to higher polynomial degrees *without* the logarithmic factor. The regularity assumptions are quite high then.

## 2.3 A posteriori error analysis

**2.3.1 Definition:** Let  $u \in V$  be the solution to a boundary value in weak form and  $u_h \in V_h$  be its finite element approximation on the mesh  $\mathbb{T}_h$ . We call a quantity  $\eta_h(u_h)$  a **a posteriori error estimator**,

$$\|u - u_h\| \leq c\eta_h(u_h). \quad (2.65)$$

The estimator is **reliable**, if the constant  $c$  is computable. It is **efficient**, if the converse estimate holds, namely

$$\eta_h(u_h) \leq c\|u - u_h\|. \quad (2.66)$$

### 2.3.1 Quasi-interpolation in $H^1$

**2.3.2.** Interpolation in Sobolev spaces in Theorem 2.2.9 relies on the nodal interpolation operator in Definition 2.2.6, which in turn requires point values of the interpolated function. Therefore, it is not defined on  $H^1(\Omega)$  in dimensions greater than one. Since we need such interpolation operators in the analysis of a posteriori error estimates, we provide them in this section.

Most details in this section are from [Verfürth, 2013]. For the ease of presentation, we present the results for Dirichlet problem of the Laplacian, namely  $V = H_0^1(\Omega)$  and

$$a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, d\mathbf{x}.$$

**2.3.3 Definition:** A shape-regular family of meshes  $\{\mathbb{T}_h\}$  is called **locally quasi-uniform**, if there is a constant  $c$  such that for every pair of cells  $T_1$  and  $T_2$  sharing at least one vertex there holds

$$h_{T_1} \leq ch_{T_2}. \quad (2.67)$$

**2.3.4 Definition:** For a vertex or higher-dimensional boundary facet  $F$ , we define the set of cells

$$\mathbb{T}_F = \{T \in \mathbb{T} \mid F \subset \partial T\}. \quad (2.68)$$

Similarly, the set of cells sharing at least one vertex with  $T$  is called  $\mathbb{T}_T$ . Additionally, we define the subdomains

$$\bar{\Omega}_F = \bigcup_{T \in \mathbb{T}_F} \bar{T}, \quad \bar{\Omega}_T = \bigcup_{T' \in \mathbb{T}_T} \bar{T}'. \quad (2.69)$$

**2.3.5 Theorem (Clément quasi-interpolation):** Let  $\{\mathbb{T}_h\}$  be a locally quasi-uniform family of meshes with piecewise polynomial finite element spaces  $V_h \subset H_0^1(\Omega)$ . Then, there exist bounded operators  $\bar{I}_h: H_0^1(\Omega) \rightarrow V_h$  such that for every function  $u \in H_0^1(\Omega)$ , every mesh cell  $T$ , every face  $F$ , and for  $m = 0, 1$  there holds

$$\|u - \bar{I}_h u\|_{m;T} \leq ch_T^{1-m} |u|_{1;\Omega_T} \quad (2.70)$$

$$\|u - \bar{I}_h u\|_{0;F} \leq ch_T^{1/2} |u|_{1;\Omega_T} \quad (2.71)$$

*Proof.* We construct the quasi-interpolation operator on a mesh cell  $T$  into the lowest order space  $\mathbb{P}_1$  or  $\mathbb{Q}_1$  in two steps. First, for each vertex  $\mathbf{X}_i$  of  $T$ , let

$$\bar{u}_i = \frac{1}{|\Omega_{\mathbf{X}_i}|} \int_{\Omega_{\mathbf{X}_i}} u \, d\mathbf{x}. \quad (2.72)$$

By Poincaré inequality,

$$\|u - \bar{u}_i\|_{\Omega_{\mathbf{X}_i}} \leq c \operatorname{diam}(\Omega_{\mathbf{X}_i}) \|\nabla u\|_{\Omega_{\mathbf{X}_i}}. \quad (2.73)$$

For a vertex  $\mathbf{X}_i$  on the boundary, we observe

$$\begin{aligned} \|\bar{u}_i\|_{\Omega_{\mathbf{X}_i}}^2 &= \frac{1}{|\Omega_{\mathbf{X}_i}|^2} \int_{\Omega_{\mathbf{X}_i}} \left( \int u \, d\mathbf{y} \right)^2 d\mathbf{x} \\ &= \frac{1}{|\Omega_{\mathbf{X}_i}|} \left( \int 1u \, d\mathbf{y} \right)^2 \\ &\leq \frac{1}{|\Omega_{\mathbf{X}_i}|} \int 1^2 d\mathbf{y} \int u^2 d\mathbf{y} \\ &= \|u\|_{\Omega_{\mathbf{X}_i}}^2 \\ &\leq c \operatorname{diam}(\Omega_{\mathbf{X}_i})^2 \|\nabla u\|_{\Omega_{\mathbf{X}_i}}^2, \end{aligned} \quad (2.74)$$

by Friedrichs' inequality, since the subdomain  $\Omega_{\mathbf{X}_i}$  has at least one face on  $\partial\Omega$ .

Now, we define the quasi-interpolation operator  $\bar{I}_h: H_0^1(\Omega) \rightarrow V_h$  cellwise on simplices by

$$\bar{I}_h u|_T = \sum_{\mathbf{x}_i \in \partial T \cap \Omega} \lambda_i \bar{u}_i. \quad (2.75)$$

Note, that zero boundary conditions are enforced by omitting vertices on the boundary. On quadrilaterals, we use the bilinear shape functions associated with the vertices instead of the barycentric coordinates  $\lambda_i$ . Now, using  $\sum \lambda_i = 1$  and  $0 \leq \lambda_i \leq 1$ , we estimate

$$\begin{aligned} \|u - \bar{I}_h u\|_T &\leq \sum_{\mathbf{x}_i \in \partial T} \|\lambda_i(u - \bar{u}_i)\|_T + \sum_{\mathbf{x}_i \in \partial T \cap \partial \Omega} \|\lambda_i \bar{u}_i\| \\ &\leq \sum_{\mathbf{x}_i \in \partial T} \|u - \bar{u}_i\|_T + \sum_{\mathbf{x}_i \in \partial T \cap \partial \Omega} \|\bar{u}_i\| \\ &\leq \text{diam}(\Omega_{\mathbf{x}_i}) \left( \sum_{\mathbf{x}_i \in \partial T} \|\nabla u\|_{\Omega_{\mathbf{x}_i}} + \sum_{\mathbf{x}_i \in \partial T \cap \partial \Omega} \|\nabla u\|_{\Omega_{\mathbf{x}_i}} \right), \end{aligned}$$

where we used (2.73) and (2.74) in the end. Observing that both sums extend over finitely many vertices and that by local quasi-uniformity we can bound the diameter of  $\text{diam}(\Omega_{\mathbf{x}_i})$  by that of  $T$ , namely  $\text{diam}(\Omega_{\mathbf{x}_i}) \leq ch_T$ , we obtain the estimate in  $L^2(T)$ . For the estimate in  $H^1(T)$ , we observe that the mean value calculation is a continuous operation on  $H_0^1(\Omega(\mathbf{x}_i))$ , and that (2.75) as a finite sum is continuous, therefore,

$$\|u - \bar{I}_h u\|_{1;T} \leq \|\nabla u\|_T + \|\nabla \bar{I}_h u\|_T \leq c \|\nabla u\|_{\Omega(T)}. \quad (2.76)$$

Finally, we use the trace estimate for  $u \in H^1(T)$

$$\|u\|_F \leq c \left( h^{-1/2} \|u\|_T + h^{1/2} \|\nabla u\|_T \right), \quad (2.77)$$

to obtain the estimate on the edge.  $\square$

**2.3.6 Theorem (Scott-Zhang quasi-interpolation):** Let  $\{\mathbb{T}_h\}$  be a locally quasi-uniform family of meshes with piecewise polynomial finite element spaces  $V_h \subset H^1(\Omega)$ . Then, there exist bounded operators  $\bar{I}_h: H^1(\Omega) \rightarrow V_h$  such that for every function  $u \in H^1(\Omega)$ , every mesh cell  $T$ , every face  $F$ , and for  $m = 0, 1$  there holds

$$\|u - \bar{I}_h u\|_{m;T} \leq ch_T^{1-m} |u|_{1;\Omega_T} \quad (2.78)$$

$$\|u - \bar{I}_h u\|_{0;F} \leq ch_T^{1/2} |u|_{1;\Omega_T}, \quad (2.79)$$

and  $\bar{I}_h u$  is a quasi-interpolation on the boundary  $\partial \Omega$ .

**2.3.7 Theorem (Schöberl quasi-interpolation):** Let  $\{\mathbb{T}_h\}$  be a locally quasi-uniform family of meshes with piecewise polynomial finite element spaces  $V_h \subset H^1(\Omega)$ . Then, there exist bounded *projection* operators  $\bar{I}_h: H^1(\Omega) \rightarrow V_h$  such that for every function  $u \in H^1(\Omega)$ , every mesh cell  $T$ , every face  $F$ , and for  $m = 0, 1$  there holds

$$\|u - \bar{I}_h u\|_{m;T} \leq ch_T^{1-m} |u|_{1;\Omega_T} \quad (2.80)$$

$$\|u - \bar{I}_h u\|_{0;F} \leq ch_T^{1/2} |u|_{1;\Omega_T}. \quad (2.81)$$

**2.3.8 Definition:** Let  $a(u, v) = f(v)$  be the weak formulation of a BVP on the space  $V$ . Then, we define the **residual**

$$\begin{aligned} R: V &\rightarrow V^* \\ w &\mapsto f(\cdot) - a(w, \cdot). \end{aligned} \quad (2.82)$$

**2.3.9 Lemma:** For  $u \in V = H_0^1(\Omega)$  solution to Poisson's equation and any other function  $w \in V$ , there holds

$$|u - v|_1 = \|Rv\|_{-1} := \sup_{w \in V} \frac{\langle Rv, w \rangle}{|w|_1}. \quad (2.83)$$

*Proof.* We have

$$\langle Rv, w \rangle = f(w) - a(v, w) = a(u - v, w). \quad (2.84)$$

Since  $a(\cdot, \cdot)$  is s.p.d., we can apply Cauchy-Schwarz to obtain the result.  $\square$

**2.3.10 Definition:** Let  $u$  be a piecewise continuous function on a mesh  $\mathbb{T}$ . On a face  $F$  between two cells  $T_1$  and  $T_2$ , we define the **mean value operator**

$$\{u\}(\mathbf{x}) = \frac{u_1(\mathbf{x}) + u_2(\mathbf{x})}{2} = \lim_{\varepsilon \searrow 0} \frac{u(\mathbf{x} - \varepsilon \mathbf{n}_1) + u(\mathbf{x} - \varepsilon \mathbf{n}_2)}{2}, \quad (2.85)$$

where  $\mathbf{n}_i$  are the outer normal vectors of the two cells. The **jump operator** is

$$\llbracket u \rrbracket = \frac{(u_1 - u_2)\mathbf{n}_1}{2} = \frac{(u_2 - u_1)\mathbf{n}_2}{2}. \quad (2.86)$$

**2.3.11 Definition:** Let  $u \in V = H_0^1(\Omega)$  be the solution to Poisson's equation and let  $u_h \in V_h$  be a finite element function. The strong form of the residual is

$$\langle Rv, w \rangle = \sum_{T \in \mathbb{T}_h} \int_T r_T(u_h) w \, d\mathbf{x} - \sum_{F \in \mathbb{F}_h^i} 2 \{ \mathbf{n} \cdot \nabla u_h \} w \, ds, \quad (2.87)$$

where

$$r_T(u_h) = f + \Delta u_h. \quad (2.88)$$

**2.3.12 Lemma:** Let  $\mathbb{T}_h$  be a locally quasi-uniform mesh. There is a constant  $c > 0$  such that the error between the true solution  $u \in V = H_0^1(\Omega)$  and the finite element solution  $u_h \in V_h \subset V$  is bounded by

$$|u - u_h|_{1; \mathbb{T}_h} \leq c \left( \sum_{T \in \mathbb{T}_h} h_T^2 \|r_T(u_h)\|_T^2 + \sum_{F \in \mathbb{F}_h^i} h_F \|\{ \mathbf{n} \cdot \nabla u_h \}\|_F^2 \right)^{1/2}. \quad (2.89)$$

*Proof.* We begin with the equivalence of error in  $V$  and residual in  $V^*$ . For the residual, there holds Galerkin orthogonality, such that we obtain

$$|u - u_h|_1 = \sup_{w \in V} \frac{\langle Ru_h, w \rangle}{|w|_1} = \sup_{w \in V} \frac{\langle Ru_h, w - \bar{I}_h w \rangle}{|w|_1}. \quad (2.90)$$

Using the strong form and the quasi-interpolation  $\bar{I}_h$ , we get

$$\begin{aligned} \langle Ru_h, w \rangle &= \sum_{T \in \mathbb{T}_h} \int_T r_T(u_h) (w - \bar{I}_h w) \, d\mathbf{x} - \sum_{F \in \mathbb{F}_h^i} 2 \{ \mathbf{n} \cdot \nabla u_h \} (w - \bar{I}_h w) \, ds \\ &\leq \sum_{T \in \mathbb{T}_h} \|r_T(u_h)\|_T \|w - \bar{I}_h w\|_T + \sum_{F \in \mathbb{F}_h^i} \|\{ \mathbf{n} \cdot \nabla u_h \}\|_T \|w - \bar{I}_h w\|_T \\ &\leq c_1 \sum_{T \in \mathbb{T}_h} h_T \|r_T(u_h)\|_T \|\nabla w\|_{\Omega_T} + c_2 \sum_{F \in \mathbb{F}_h^i} h_E^{1/2} \|\{ \mathbf{n} \cdot \nabla u_h \}\|_T \|\nabla w\|_{\Omega_T} \end{aligned}$$

Applying Hölder inequality, we obtain for the first term

$$\sum_{T \in \mathbb{T}_h} h_T \|r_T(u_h)\|_T \|\nabla w\|_{\Omega_T} \leq \left( \sum_{T \in \mathbb{T}_h} h_T^2 \|r_T(u_h)\|_T^2 \right)^{1/2} \left( \sum_{T \in \mathbb{T}_h} \|\nabla w\|_{\Omega_T}^2 \right)^{1/2},$$

and similar for the second term. Both contain a term of the form

$$\sum_{T \in \mathbb{T}_h} \|\nabla w\|_{\Omega_T}^2 = \sum_{T \in \mathbb{T}_h} \sum_{T' \in \mathbb{T}_T} \|\nabla w\|_{T'}^2 \leq n \|\nabla w\|_{\Omega}^2,$$

where  $n$  is the maximal number of occurrences of a cell  $T'$  in the double sum. This  $n$  is bounded uniformly on shape regular family of meshes, since shape regularity prohibits degeneration of cells. Therefore, we conclude

$$\langle Ru_h, w \rangle \leq c \left( \sum_{T \in \mathbb{T}_h} h_T^2 \|r_T(u_h)\|_T^2 + \sum_{F \in \mathbb{F}_h^i} h_F \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F^2 \right)^{1/2} |w|_{1;\Omega}.$$

Entering into (2.90) yields the proposition.  $\square$

**Remark 2.3.13.** The constant  $c$  in the previous theorem depends on the constant in the quasi-interpolation estimate, which in turn was derived using Poincaré and trace inequalities. Traditionally, both are derived using indirect arguments, but a constructive proof with computable bounds is possible. For details, see [Verfürth, 2013, Chapter 3]. There still remains the question, whether these bounds are sufficiently sharp to be applicable in practice.

**2.3.14 Definition:** For a function  $f \in L^2(\Omega)$ , we define the cell-wise average

$$\bar{f}(\mathbf{x}) = \frac{1}{|T|} \int_T f \, d\mathbf{x} \quad \mathbf{x} \in T, \quad T \in \mathbb{T}_h, \quad (2.91)$$

and the **data oscillation**

$$\text{osc}_T f = \|f - \bar{f}\|_T. \quad (2.92)$$

**2.3.15 Definition:** The residual based error estimator for the finite element method for Poisson's equation on a mesh  $\mathbb{T}$  is defined as

$$\eta_h(u_h) = \sum_{T \in \mathbb{T}} \eta_T(u_h), \quad (2.93)$$

where

$$\eta_T(u_h)^2 = h_T^2 \|\bar{f} + \Delta u_h\|_T^2 + \frac{1}{2} \sum_{F \subset \partial T} h_F \|\llbracket n \cdot \nabla u_h \rrbracket\|_F^2. \quad (2.94)$$



**2.3.16 Theorem:** There holds with positive constants  $c_1$  and  $c_2$

$$c_1 |u - u_h|_1^2 \leq \eta_h(u_h)^2 + \sum_{T \in \mathbb{T}_h} h_T^2 \text{osc}_T^2 f \quad (2.95)$$

$$c_2 \eta_T^2(u_h) \leq |u - u_h|_{1;\Omega_T}^2 + \sum_{T \in \mathbb{T}_T} h_T^2 \text{osc}_T^2 f \quad (2.96)$$

*Proof.* Note that the right hand side of the estimate (2.89) in Lemma 2.3.12 differs from the estimator  $\eta_h(u_h)$  only by the replacement of  $f$  by  $\bar{f}$ . Therefore, we can start the proof of this lemma by changing (2.90) to

$$|u - u_h|_1 = \sup_{w \in V} \frac{\langle Ru_h, w \rangle}{|w|_1} = \sup_{w \in V} \frac{\langle \bar{R}u_h, w \rangle + (f - \bar{f}, w)}{|w|_1}, \quad (2.97)$$

where we ad hoc defined  $\bar{R}$  like  $R$ , but replacing  $r_T = f + \nabla u_h$  by  $\bar{f} + \nabla u_h$ . Due to equality, we can still subtract the quasi-interpolant of  $w$  and continue through the whole proof. What remains is the estimate

$$(f - \bar{f}, w - \bar{I}_h w) \leq c \sum_{T \in \mathbb{T}} h_T \|f - \bar{f}\|_T \|\nabla w\|_{\Omega_T} \leq c \left( \sum_{T \in \mathbb{T}} h_T \|f - \bar{f}\|_T \right)^{1/2} |w|_{1;\Omega}, \quad (2.98)$$

which is proven by Hölder inequality and the boundedness of the number of cells in  $\mathbb{T}_T$ . Thus, the upper bound (2.95) is an immediate consequence of Lemma 2.3.12.

Note: the proof for the lower bound is for linear elements only. It can be generalized to higher order by generalizing the bubble functions below, which is mostly technical.

We observe that the lower bound is local, that is, the estimate on each cell  $T$  is bounded from above by the error on the halo  $\Omega_T$ . Therefore, we start by constructing test functions with support on a cell. On a simplex, we define the **bubble function** in terms of barycentric coordinates

$$B_T = \frac{1}{(d+1)^{d+1}} \lambda_0 \lambda_1 \dots \lambda_d, \quad (2.99)$$

while on the reference hypercube  $(0, 1)^d$  we let

$$\hat{B}_T(\hat{\mathbf{x}}) = b(\hat{x}_1) b(\hat{x}_2) \dots b(\hat{x}_d) \quad b(x) = 4x(1-x). \quad (2.100)$$

Both share the following properties: they vanish on  $\partial T$ , they are positive inside  $T$ , and  $\max B_T = 1$ . From positivity, we deduce the existence of a constant  $c$  depending on shape regularity and the shape function space  $\mathcal{P}(T)$ , such that

$$\int_T p^2 B_T \, d\mathbf{x} \geq c \|p\|_T^2 \quad \forall p \in \mathcal{P}(T). \quad (2.101)$$

Furthermore,  $B_T$  is polynomial, such that the inverse estimate holds.

Choose now the test function  $w_T = (\bar{f} + \Delta u_h)B_T$ . Since it vanishes on the boundary and outside of  $T$ , the strong and weak form of the residual reduce to

$$\int_T r_T(u_h)w_T \, d\mathbf{x} = \int_T \nabla(u - u_h) \cdot \nabla w_T \, d\mathbf{x}. \quad (2.102)$$

Adding the difference of  $f$  and  $\bar{f}$  on both sides yields

$$\int_T (\bar{f} + \Delta u_h)^2 B_T \, d\mathbf{x} = \int_T \nabla(u - u_h) \cdot \nabla w_T \, d\mathbf{x} + \int_T (\bar{f} - f)w_T \, d\mathbf{x}. \quad (2.103)$$

On the left, we estimate

$$c\|\bar{f} + \Delta u_h\|_T^2 \leq \int_T (\bar{f} + \Delta u_h)^2 B_T \, d\mathbf{x}. \quad (2.104)$$

On the right, we estimate the residual term by

$$\begin{aligned} \int_T \nabla(u - u_h) \cdot \nabla w_T \, d\mathbf{x} &\leq |u - u_h|_{1;T} |w_T|_{1;T} \\ &\leq c|u - u_h|_{1;T} h_T^{-1} \|\bar{f} + \Delta u_h\|_T. \end{aligned}$$

The data oscillation is estimated in a straightforward way by

$$\int_T (\bar{f} - f)w_T \, d\mathbf{x} \leq \text{osc}_T f \|\bar{f} + \Delta u_h\|_T$$

Combining these three estimates yields

$$c\|\bar{f} + \Delta u_h\|_T \leq h_T^{-1}|u - u_h|_{1;T} + \text{osc}_T f, \quad (2.105)$$

which is the desired estimate for the cell term of the estimator  $\eta_T(u_h)$  if we multiply by  $h_T$ .

The estimate for the face term is similar, using a bubble function for the face. Since this function is nonzero on the face, its support extends over two mesh cells. On simplicial cells, it is constructed like the cell bubble in equation (2.99), but extending the product only over indices of vertices on the face. Therefore, it is a polynomial of degree  $d - 1$  on the face and also in the interior of each cell. For hypercubes, it is the product of the quadratic polynomials  $b(x)$  of variables on the face times a linear function decaying linearly from 1 to zero. Again, we normalize such that the maximum equals 1. Choosing<sup>1</sup>  $w_F = 2\{\{\mathbf{n} \cdot \nabla u_h\}\}_F B_F$ ,

---

<sup>1</sup>For linear elements, the derivative on the face is constant and thus this definition is obvious. For higher order polynomials, we need an extension from the face into the interior, which can be obtained from the barycentric coordinates or the tensor product structure.

we obtain

$$\begin{aligned}
c \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F^2 &\leq \int_F \llbracket \mathbf{n} \cdot \nabla u_h \rrbracket^2 B_F \, ds \\
&= \int_F \llbracket \mathbf{n} \cdot \nabla u_h \rrbracket w_F \, ds \\
&= \sum_{T \in \mathbb{T}_F} \int_T r_T w_F \, d\mathbf{x} - \langle R, w_F \rangle \\
&= \sum_{T \in \mathbb{T}_F} \int_T r_T w_F \, d\mathbf{x} - \int_{\Omega_F} \nabla(u - u_h) \cdot \nabla w_F \, d\mathbf{x} \\
&= \sum_{T \in \mathbb{T}_F} \int_T (\bar{f} + \Delta u_h) w_F \, d\mathbf{x} - \int_{\Omega_F} \nabla(u - u_h) \cdot \nabla w_F \, d\mathbf{x} + \sum_{T \in \mathbb{T}_F} \int_T (\bar{f} - f) w_F \, d\mathbf{x}.
\end{aligned}$$

Again, we estimate the three terms on the right hand side separately. For the first, we estimate the norm of  $w_F$  by its trace on the edge  $F$ . From the already proven estimate for the cell residual (2.105), we obtain

$$\begin{aligned}
\sum_{T \in \mathbb{T}_F} \int_T (\bar{f} + \Delta u_h) w_F \, d\mathbf{x} &\leq \sum_{T \in \mathbb{T}_F} \|\bar{f} + \Delta u_h\|_T \|w_F\|_T \\
&\leq \sum_{T \in \mathbb{T}_F} \|\bar{f} + \Delta u_h\|_T c h_F^{1/2} \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F \\
&\leq +c \sum_{T \in \mathbb{T}_F} \left( h_F^{-1/2} |u - u_h|_{1;T} + h_F^{1/2} \text{osc}_T f \right) \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F
\end{aligned}$$

Similarly, the inverse estimate for  $w_F$  yields

$$\begin{aligned}
\int_{\Omega_F} \nabla(u - u_h) \cdot \nabla w_F \, d\mathbf{x} &\leq |u - u_h|_{1;\Omega_F} \|\nabla w_F\|_{\Omega_F} \\
&\leq |u - u_h|_{1;\Omega_F} c h_F^{-1/2} \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F.
\end{aligned}$$

Finally, the data oscillation terms becomes

$$\begin{aligned}
\sum_{T \in \mathbb{T}_F} \int_T (\bar{f} - f) w_F \, d\mathbf{x} &\leq \sum_{T \in \mathbb{T}_F} \|\bar{f} - f\|_T \|w_F\|_T \\
&\leq \sum_{T \in \mathbb{T}_F} \|\bar{f} - f\|_T c h_F^{1/2} \|\llbracket \mathbf{n} \cdot \nabla u_h \rrbracket\|_F.
\end{aligned}$$

Summing and scaling with  $h_F^{1/2}$  yields the result.  $\square$

## Chapter 3

# Variational Crimes

**3.0.1.** In the previous chapter, we considered finite element methods applying the original bilinear form to a subspace  $V_h \subset V$ . This assumes, that the domain  $\Omega$  is exactly represented by the mesh, and that all integrals are computed exactly. Both assumptions reduce the applicability of the finite element method considerably. Therefore, we now extend our analysis to cases, where we allow  $V_h \not\subset V$  and discrete bilinear forms  $a_h(.,.) \neq a(.,.)$ .

### 3.1 Numerical quadrature

**3.1.1.** We begin the investigation of variational crimes by studying the effect of using numerical quadrature instead of exact integration on mesh cells. In particular, we investigate approximations of the form

$$\int_T f(\mathbf{x}) \, d\mathbf{x} \approx \sum_{k=1}^{n_q} \omega_k f(\mathbf{x}_k) =: Q_T(f). \quad (3.1)$$

First, we observe that  $Q_T$  is not a bounded operator on  $L^1(\Omega)$ , such that  $Q_T(\nabla u \cdot \nabla v)$  is undefined for functions in  $H^1(\Omega)$ . The surprising result of this section is, that quadrature is still admissible for the implementation of a finite element method.

We will first set a theoretical framework and then investigate quadrature rules in detail. The presentation follows [Ciarlet, 1978, Chapter 4].

**3.1.2 Lemma (Strang's first lemma):** Let  $a(.,.)$  be a bounded and elliptic bilinear form on the Hilbert space  $V$ . Let  $V_n \subset V$  and let  $a_h(.,.)$  be a bilinear form, bounded and elliptic on  $V_n$  with constants  $M_n$  and  $\alpha_n$ . Let  $f, f_n \in V^*$ . If  $u \in V$  and  $u_n \in V_n \subset V$  are solutions to

$$\begin{aligned} a(u, v) &= f(v) & \forall v \in V, \\ a_n(u_n, v_n) &= f_n(v_n) & \forall v_n \in V_n, \end{aligned}$$

respectively, there holds

$$\begin{aligned} \|u - u_n\|_V &\leq \inf_{v_n \in V_n} \left[ \left(1 + \frac{M}{\alpha_n}\right) \|u - v_n\|_V + \frac{1}{\alpha_n} \|a_n(v_n, \cdot) - a(v_n, \cdot)\|_{V_h^*} \right] \\ &\quad + \frac{1}{\alpha_n} \|f_n - f\|_{V_h^*}. \end{aligned} \quad (3.2)$$

*Proof.* Using  $V_n$ -ellipticity of  $a_n(.,.)$  yields for arbitrary  $v_n \in V_n$

$$\begin{aligned} \alpha_n \|u_n - v_n\|^2 &\leq a_n(u_n - v_n, u_n - v_n) \\ &= f_n(u_n - v_n) + a(u - v_n, u_n - v_n) - f(u_n - v_n) \\ &\quad + a(v_n, u_n - v_n) - a_n(v_n, u_n - v_n) \end{aligned}$$

We estimate separately

$$\begin{aligned} \frac{a(u - v_n, u_n - v_n)}{\|u_n - v_n\|} &\leq M \|u_n - v_n\|, \\ \frac{|f_n(u_n - v_n) - f(u_n - v_n)|}{\|u_n - v_n\|} &\leq \sup_{w_n \in V_n} \frac{|f_n(w_n) - f(w_n)|}{\|w_n\|}, \\ \frac{|a(v_n, u_n - v_n) - a_n(v_n, u_n - v_n)|}{\|u_n - v_n\|} &\leq \sup_{w_n \in V_n} \frac{|a_n(v_n, w_n) - a(v_n, w_n)|}{\|w_n\|}. \end{aligned}$$

Combining all terms, we obtain

$$\begin{aligned} \|u - u_n\| &\leq \|u - v_n\| + \|u_n - v_n\| \\ &\leq \|u - v_n\| + \frac{1}{\alpha_n} (M \|u - v_n\| + \|a_n(v_n, \cdot) - a(v_n, \cdot)\|_{V_h^*} + \|f_n - f\|_{V_h^*}). \end{aligned} \quad (3.3)$$

□

**Remark 3.1.3.** Strang's lemma states, that the error can be split into the approximation error of the space  $V_h$  and the consistency error of the approximations  $a_h(.,.)$  and  $f_h(.,.)$ . Both errors are scaled with the stability factor  $1/\alpha_n$

of the discrete problem. So far, this is consistent with error estimates for instance for Runge-Kutta methods. There is an important difference though: the consistency error is not evaluated for the exact solution, but only for its discrete approximation.

**Remark 3.1.4.** We will apply Strang's lemma to a family of meshes indexed by mesh size  $h$  and assess the infimum by an interpolation operator. It is clear, that we will only obtain optimal convergence rates compared to the interpolation estimate, if there exists  $\alpha_0 > 0$  such that  $\alpha_n \geq \alpha_0$  uniformly with respect to  $n$ . While it is not a prerequisite of Strang's lemma, it is our goal for all discretizations.

**Remark 3.1.5.** Quadrature would be infeasible, if we had to devise a quadrature rule for every mesh cell  $T$ . Instead, we tabulate quadrature formulas for the reference cell  $\hat{T}$  by choosing quadrature points  $\hat{\mathbf{x}}_k$  and weights  $\omega_k$  and write

$$Q_{\hat{T}}(\hat{f}) = \sum_{k=1}^{n_q} \omega_k \hat{f}(\hat{\mathbf{x}}_k). \quad (3.4)$$

We compute integrals over  $T$  though mapping,

$$Q_T(f) = \sum_{k=1}^{n_q} \det \nabla \Phi_T(\hat{\mathbf{x}}_k) \omega_k \hat{f}(\hat{\mathbf{x}}_k). \quad (3.5)$$

Thus,  $Q_T$  is defined by quadrature points  $\mathbf{x}_k = \Phi(\hat{\mathbf{x}}_k)$  and quadrature weights  $\det \nabla \Phi_T(\hat{\mathbf{x}}_k) \omega_k$ .

Quadrature rules on the reference cell  $\hat{T}$  are obtained by interpolation after choosing quadrature points by employing the properties of orthogonal polynomials. The construction of such quadrature rules for simplices is somewhat complicated, and we refer to tables in the cited literature.

An important consequence of the use of quadrature points as roots of orthogonal polynomials is the fact that all weights are positive.

**3.1.6 Definition:** Given a one-dimensional quadrature rule  $Q_I$  on the interval  $I = [0, 1]$  with

$$Q_I(\hat{f}) = \sum_{k=1}^{n_q} \omega_k \hat{f}(\hat{x}_k). \quad (3.6)$$

Then, a quadrature rule on  $\hat{T} = [0, 1]^d$  is defined by

$$Q_{\hat{T}}(\hat{f}) = \sum_{k_1=1}^{n_q} \cdots \sum_{k_d=1}^{n_q} \omega_{k_1} \cdots \omega_{k_d} \hat{f}(\hat{x}_{k_1}, \dots, \hat{x}_{k_d}). \quad (3.7)$$

**3.1.7 Lemma:** If  $Q_I$  is an  $n$ -point Gauß formula, the tensor product quadrature in the preceding definition is exact on  $\mathbb{Q}_{2n-1}$ .

**Remark 3.1.8.** When we look at Poisson's equation on simplicial meshes, the integrals

$$a_T(\varphi_j, \varphi_i) = \int_T \nabla \varphi_j \cdot \nabla \varphi_i \, d\mathbf{x} = \int_{\hat{T}} \nabla \Phi^{-T} \hat{\nabla} \hat{\varphi}_j \cdot \nabla \Phi^{-T} \hat{\varphi}_i \det \nabla \Phi \, d\hat{\mathbf{x}}, \quad (3.8)$$

can be computed exactly, since  $\nabla \Phi$  is a constant matrix. Already on arbitrary quadrilaterals,  $\nabla \Phi^{-T}$  is a rational function, which is harder to integrate. But, in order to justify the use and investigate the properties of approximations by numerical quadrature, we are considering a model problem with varying coefficients.

**3.1.9 Assumption:** Let in this section the bilinear form be defined as

$$a(u, v) = \int_{\Omega} \nabla u A \nabla v^T \, dx, \quad (3.9)$$

where  $A = A(\mathbf{x})$  is a matrix with

$$\xi^T A(\mathbf{x}) \xi \geq \alpha |\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \quad \forall \mathbf{x} \in \Omega. \quad (3.10)$$

The function  $u \in V = H_0^1(\Omega)$  is the solution to

$$a(u, v) = (f, v) \quad \forall v \in V. \quad (3.11)$$

We assume that the coefficients  $a_{ij}$  and the right hand side  $f$  are functions which are well-defined in all quadrature points on all meshes of a family  $\{\mathbb{T}_h\}$ .

**3.1.10 Definition:** The bilinear form  $a_h(.,.)$  obtained by numerical quadrature  $Q_{\hat{T}}$  on the mesh  $\mathbb{T}_h$  is defined as

$$a_h(u, v) = \sum_{T \in \mathbb{T}_h} Q_T(\nabla u A \nabla v^T), \quad (3.12)$$

where  $Q_T$  is obtained from  $Q_{\hat{T}}$  by the mapping  $\Phi_T$ . The right hand side is

$$f_h(v) = \sum_{T \in \mathbb{T}_h} Q_T(fv). \quad (3.13)$$

**3.1.11 Lemma:** Let a finite element method be defined by a mesh  $\mathbb{T}_h$  and a shape function space  $\mathcal{P}$  on the reference cell  $\hat{T}$ . Let  $Q_{\hat{T}}$  be a quadrature formula, such that all quadrature weights are positive. Furthermore, assume that one of the following holds:

1. The function values in the quadrature points  $\hat{\mathbf{x}}_k$  are unisolvent on

$$\mathcal{G}_1 = \{\partial_i p | p \in \mathcal{P}, i = 1, \dots, d\}, \quad (3.14)$$

2. The quadrature rule is exact on

$$\mathcal{G}_2 = \{\nabla p \cdot \nabla q | p, q \in \mathcal{P}\}. \quad (3.15)$$

Then, there is a constant  $\alpha_0 > 0$  depending on shape regularity and the quadrature rule, but independent of  $h$ , such that

$$a_h(u_h, u_h) \geq \alpha_0 |u_h|_{1;\Omega}^2 \quad \forall u_h \in V_h. \quad (3.16)$$

**Remark 3.1.12.** If  $\mathcal{P} = \mathbb{P}_k$ , then  $\mathcal{G}_1 = \mathbb{P}_{k-1}$  and  $\mathcal{G}_2 = \mathbb{P}_{2k-2}$ . For  $\mathbb{Q}_k$ , these relations are more complicated, and the simplest we can say is  $\mathcal{G}_1 \subset \mathbb{Q}_k$  and  $\mathcal{G}_2 \subset \mathbb{Q}_{2k}$ .

*Proof of Lemma 3.1.11.* Assume first condition 1. From the positivity of quadrature weights, we conclude that

$$Q_{\hat{T}}(\nabla p \cdot \nabla p) = 0 = \sum_{k=1}^{n_q} \sum_{i=1}^d \omega_k (\partial_i p(\hat{\mathbf{x}}_k))^2 = 0 \quad (3.17)$$

implies  $\partial_i p(\hat{\mathbf{x}}_k) = 0$ . Unisolvence implies  $\partial_i p \equiv 0$ . Therefore,  $\sqrt{Q_{\hat{T}}(\nabla p \cdot \nabla p)}$  defines a norm on the space  $\mathcal{P}/\mathbb{R}$ . Since this space is finite dimensional and  $|\cdot|_{1;\hat{T}}$  is a second norm, we conclude from norm equivalence the existence of a constant  $c > 0$  such that

$$c|p|_{1;\hat{T}}^2 \leq Q_{\hat{T}}(\nabla p \cdot \nabla p) \quad \forall p \in \mathcal{P}. \quad (3.18)$$

On the other hand, condition 2 yields the same estimate with  $c = 1$ .

It is easily verified, that  $|p|_{1;\hat{T}}^2$  and  $Q_{\hat{T}}(\nabla p \cdot \nabla p)$  scale equally under the mapping  $\Phi_T$ , such that the same estimate holds on  $T$  with a different constant depending on shape regularity, again denoted by  $c$ . Finally,

$$\begin{aligned} \alpha_0 c |u_h|_{1;\Omega}^2 &= \alpha_0 c \sum_{T \in \mathbb{T}_h} |u_h|_{1;T}^2 \\ &\leq \alpha_0 \sum_{T \in \mathbb{T}_h} Q_{\hat{T}}(\nabla u_h \cdot \nabla u_h) \\ &\leq \sum_{T \in \mathbb{T}_h} Q_{\hat{T}}(\nabla u_h A \nabla u_h^T) = a_h(u_h, u_h). \end{aligned}$$



□

**3.1.13.** Now that we have established conditions for stability of the discrete problem, we can address an estimate of the consistency error. In order to avoid additional overhead in an already technical argument, we restrict the analysis to the spaces  $\mathcal{P}_T = \mathbb{P}_k$  on simplices. Instead of proving a general result for arbitrary quadrature formulas, we first take ask, what the order of the quadrature error should be. Indeed, we have the interpolation estimate  $\|u - I_h u\|_1 = \mathcal{O}(h^k)$ . Therefore, we investigate quadrature rules introducing consistency errors of the same order.

**3.1.14 Theorem:** Assume in addition to Assumption 3.1.9 for  $k \geq 1$  that for all cells  $\mathcal{P}_T = \mathbb{P}_k$  and  $a_{ij} \in W^{k,\infty}(\Omega)$ . Let the quadrature rule  $Q_{\hat{T}}$  be exact for polynomials in  $\mathbb{P}_{2k-2}$ . Then, there exists a constant  $c$  such that for all meshes of a quasi-uniform family  $\{\mathbb{T}_h\}$  of meshes there holds

$$|a(v_h, w_h) - a_h(v_h, w_h)| \leq ch^k \|A\|_{k,\infty;\Omega} \|v_h\|_{k;h} |w_h|_{1;h}. \quad (3.19)$$

**3.1.15 Lemma:** Let  $u \in W^{k,p}(\Omega)$  and  $v \in W^{k,\infty}(\Omega)$ . Then,  $uv \in W^{k,p}(\Omega)$  and

$$|uv|_{k,p;\Omega} \leq c \sum_{i=0}^k |u|_{i,q;\Omega} |v|_{k-i,\infty;\Omega}, \quad (3.20)$$

with a constant  $c$  depending on  $k$  and the space dimension, but not on  $\Omega$ .

*Proof.* This is basically the product rule. For a multi-index  $\alpha$  with  $|\alpha| = k$ , we have

$$\partial^\alpha(uv) = \sum_{i=0}^k \sum_{\substack{|\beta|=i \\ \beta_j \leq \alpha_j}} \partial^\beta u \partial^{\alpha-\beta} v.$$

Then, Hölder inequality is applied to the sums and the integrals. □

*Proof of Theorem 3.1.14.* We first argue on the reference cell and define the quadrature error

$$E_{\hat{T}}(f) = \int_{\hat{T}} f \, d\hat{\mathbf{x}} - Q_{\hat{T}}(f). \quad (3.21)$$

Our goal is the estimation of  $E_{\hat{T}}(apq)$  for functions  $a \in W^{k,\infty}(\hat{T})$  and polynomials  $p, q \in \mathbb{P}_{k-1}$ , such that we obtain an estimate for the constituents of  $E_{\hat{T}}(\hat{\nabla}\hat{u}\hat{A}\hat{\nabla}\hat{v})$ .

We combine  $\varphi = ap \in W^{k,\infty}(\hat{T})$ . Since  $W^{k,\infty}(\hat{T}) \hookrightarrow C(\hat{T})$ , we obtain

$$|E_{\hat{T}}(\varphi q)| \leq c\|\varphi q\|_{\infty} \leq c\|\varphi\|_{0,\infty}\|q\|_{0,\infty} \leq c\|\varphi\|_{k,\infty}\|q\|_{L^2(\hat{T})}. \quad (3.22)$$

For the last inequality, we used the natural bound of the norm in  $W^{0,\infty}$  by that of  $W^{k,\infty}$  and the norm equivalence on  $\mathbb{P}_{k-1}$ . We conclude that  $J_q(\varphi) = E_{\hat{T}}(\varphi q)$  is a bounded linear functional on  $W^{k,\infty}$ . Furthermore, it vanishes for  $\varphi \in \mathbb{P}_{k-1}$  by the assumption on the quadrature rule. Thus by the Bramble-Hilbert lemma (Lemma 2.2.3),

$$|E_{\hat{T}}(\varphi q)| \leq c|\varphi|_{k,\infty}\|w\|_{L^2(\hat{T})}. \quad (3.23)$$

Now, we apply Lemma 3.1.15, to obtain

$$\begin{aligned} |\varphi|_{k,\infty} &= |ap|_{k,\infty} \\ &\leq c \sum_{i=0}^{k-1} |a|_{k-i,\infty} |p|_{i,\infty} \\ &\leq c \sum_{i=0}^{k-1} |a|_{k-i,\infty} |p|_i, \end{aligned} \quad (3.24)$$

where we used that  $|p|_k = 0$  and norm equivalence on  $\mathbb{P}_{k-1}$ . Collecting everything and reintroducing  $\hat{a}$ ,  $\hat{p}$ , and  $\hat{q}$  for  $a$ ,  $p$ , and  $q$ , respectively, we obtain

$$|E_{\hat{T}}(\hat{a}\hat{p}\hat{q})| \leq c \left( \sum_{i=1}^{k-1} |\hat{a}|_{k-1,\infty;\hat{T}} |\hat{p}|_{i;\hat{T}} \right) \|\hat{q}\|_{L^2(\hat{T})}. \quad (3.25)$$

The scaling lemma yields

$$|\hat{a}|_{k-1,\infty;\hat{T}} \leq ch_T^{k-i} |a|_{k-i,\infty;T} \quad 0 \leq i \leq k-1, \quad (3.26)$$

$$|\hat{p}|_{i;\hat{T}} \leq ch_T^{i-d/2} |p|_{i;T} \quad 0 \leq i \leq k-1, \quad (3.27)$$

$$\|\hat{q}\|_{L^2(\hat{T})} \leq ch_T^{-d/2} \|q\|_{L^2(T)}. \quad (3.28)$$

Therefore,

$$\begin{aligned} |E_T(apq)| &\leq ch_T^{d/2} |E_{\hat{T}}(\hat{a}\hat{p}\hat{q})| \\ &\leq ch_T^k \left( \sum_{i=1}^{k-1} |a|_{k-i,\infty;T} |p|_{i;T} \right) \|q\|_{L^2(T)} \\ &\leq ch_T^k \|a\|_{k-i,\infty;T} \|p\|_{i;T} \|q\|_{L^2(T)}. \end{aligned} \quad (3.29)$$

Entering the derivatives of  $u_h$  and  $v_h$  for  $p$  and  $q$ , respectively, and summing up over all cells yields the result.  $\square$

**Remark 3.1.16.** Note that the assumption of quasi-uniformity is convenient for notation, but excessive. Indeed, shape regularity is sufficient, but in that case, the estimate must be localized, that is,

$$a(v_h, w_h) - a_h(v_h, w_h) \leq c \sum_{T \in \mathbb{T}_h} h_T^k \|A\|_{k, \infty; T} \|v_h\|_{k; T} |w_h|_{1; T}. \quad (3.30)$$

**3.1.17 Theorem:** Let the finite element use shape function spaces  $\mathcal{P}_T = \mathbb{P}_k$  and let the quadrature rule  $Q_{\hat{T}}$  be exact for polynomials in  $\mathbb{P}_{2k-2}$ . Then, there exists a constant  $c$  such that for all cells of a quasi-uniform family  $\{\mathbb{T}_h\}$  of meshes there holds

$$|E_T(fp)| \leq ch_T^k \|f\|_{k; T} \|p\|_{1; T} \quad \forall f \in H^k(\Omega), \quad \forall p \in \mathcal{P}. \quad (3.31)$$

## Chapter 4

# Solving the Discrete Problem

**4.0.1.** The motivation for the use of iterative methods lies in the fact that matrices resulting from the discretization of partial differential equations tend to be very big, but sparse, that is, most of their entries are zero. Let us take for example trilinear finite elements on a uniform grid of  $(n - 1)$  cells in each direction. This leads to  $n^3$  degrees of freedom which we number lexicographically. The matrix stencil, that is, the distribution of nonzero entries, in one dimension is tridiagonal, that is,

$$S_1 = \begin{pmatrix} * & * & & \\ * & * & * & \\ & \ddots & \ddots & \ddots \\ & & * & * \end{pmatrix}$$

In two dimensions, we obtain the stencil

$$S_2 = S_1 \otimes S_1 = \begin{pmatrix} S_1 & S_1 & & \\ S_1 & S_1 & S_1 & \\ & \ddots & \ddots & \ddots \\ & & S_1 & S_1 \end{pmatrix},$$

and in three dimensions

$$S_3 = S_1 \otimes S_1 \otimes S_1 = \begin{pmatrix} S_2 & S_2 & & \\ S_2 & S_2 & S_2 & \\ & \ddots & \ddots & \ddots \\ & & S_2 & S_2 \end{pmatrix}.$$

The corresponding matrix has  $N = n^3$  rows and columns. Let us compare some solution methods for  $n = 100$  with respect to memory and a hypothetical hardware which executes  $10^{10}$  multiplications per second (additions are free).

**Gaussian elimination** The effort needed is  $\frac{1}{3}N^3 + \mathcal{O}(N^2)$ , leading to the computing time

$$T_G = \frac{10^{18}}{3 \cdot 10^{10}} \text{ sec} \approx 3 \cdot 10^7 \text{ sec} \approx 1.06 \text{ years}$$

The backward substitution is only of order  $N^2$  and can be neglected. The memory requirement with double precision is  $8 \cdot 10^{12}$  Bytes, almost 10 terabytes.

**Banded LU decomposition** Here we make use of the fact that LU decomposition can be restricted to the hull of the outermost nonzero elements of the matrix, the so called banded or skyline version. Let  $M$  be the greatest distance of a nonzero matrix element  $a_{ij}$  from the diagonal, that is,  $M = |i - j|$ . Then, the leading term of the effort is  $\frac{1}{3}N \cdot M^2$ , yielding with  $M = n^2$  a computing time of

$$T_{BLU} = \frac{10^6 \cdot 10^4 \cdot 10^4}{3 \cdot 10^{10}} \text{ sec} \approx 3 \cdot 10^3 \text{ sec} \approx 50 \text{ minutes}$$

The storage requirement for  $N \cdot M$  double precision numbers is  $8 \cdot 10^{10}$  Bytes, almost 100 gigabytes.

**matrix vector product** For comparison, the multiplication with such a matrix, given that there are at most 9 nonzero entries per row costs

$$T_{mult} = \frac{9 \cdot 10^6}{10^{10}} \approx 1 \text{ msec},$$

that is, we can perform more than  $10^6$  matrix-vector multiplications before we reach the effort of the banded LU decomposition. The storage requirement is roughly 1 gigabyte and can be reduced to almost zero by a smart implementation.

**Remark 4.0.2.** For purposes of analysis we typically choose the space  $X = L^2(\Omega)$ . We admit a small inaccuracy here: when we run the algorithms on a computer, we usually employ the Euclidean inner product, thus  $X$  should be the space of degrees of freedom. But this is a discrete space, where we cannot use theory of function spaces easily. Instead, we note that the  $L^2$ -inner product of standard finite element bases yield inner products equivalent to the Euclidean up to the local mesh size (see Lemma ??).

**Example 4.0.3.** While the methods developed in this chapter are fairly general, we introduce a specific model problem as a simple benchmark case. To this end, we consider the Dirichlet problem: find  $u \in V = H_0^1(\Omega)$  such that

$$a(u, v) \equiv \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \equiv f(v), \quad \forall v \in V. \quad (4.1)$$

The finite dimensional linear systems of equations are derived from finite element discretizations on quasi-uniform meshes of cells with maximal diameter  $h$ ,

yielding a sequence of spaces  $V_h$ , on which linear systems are introduced by the same weak form (4.1).

**Notation 4.0.4.** With a bilinear form  $a(.,.)$  on  $X \times X$  we associate the operator  $A : X \rightarrow X$  by

$$\langle Au, v \rangle = a(u, v), \quad \forall v \in V, \quad (4.2)$$

where now  $V = \mathcal{D}(A)$  is the **domain** of  $A$ , that is, the subset of functions  $v \in X$ , such that  $Av$  is defined and in  $X$ .

We will tacitly assume that operators  $A$ ,  $B$ , etc. are defined by equation (4.2) and the bilinear forms  $a(.,.)$ ,  $b(.,.)$ , etc., respectively, if they are not defined otherwise.

**4.0.5 Definition:** We call the bilinear form  $a(.,.)$  and its associated operator  $A$  **symmetric**, if there holds

$$a(u, v) = a(v, u) \quad \forall u, v \in V.$$

They are called  **$V$ -elliptic**, if for there is a positive number  $\gamma$  such that

$$a(u, u) \geq \gamma \|u\|_V^2 \quad \forall u \in V.$$

**4.0.6 Definition:** For positive definite, symmetric operators, we obtain the possibly infinite bounds of the spectrum

$$\Lambda(A) = \sup_{u \in V} \frac{a(u, u)}{\|u\|_X^2}, \quad \lambda(A) = \inf_{u \in V} \frac{a(u, u)}{\|u\|_X^2}, \quad (4.3)$$

as well as the possibly infinite **spectral condition number**

$$\kappa(A) = \frac{\Lambda(A)}{\lambda(A)}.$$

**Remark 4.0.7.** Note that the spectral condition number depends on the norm of the space  $X$ . It is bounded, if and only if  $A$  is bounded with respect to this norm.

**Example 4.0.8.** Let  $X = H_0^1(\Omega)$  with the inner product

$$\langle u, v \rangle_1 = \int_{\Omega} \nabla u \cdot \nabla v \, dx.$$

If  $A$  is the operator associated with the bilinear form  $a(.,.)$  in (4.1), then

$$\Lambda(A) = \lambda(A) = \kappa(A) = 1.$$

If on the other hand  $X = L^2(\Omega)$  equipped with the usual  $L^2$ -inner product, then  $A$  is unbounded and thus  $\kappa(A) = \infty$ .  $\lambda(A)$  is the constant in Friedrichs' inequality.

**Notation 4.0.9.** After choosing a basis for a finite dimensional space  $X_n$  or a Schauder basis for the space  $X$  (assuming  $X$  separable), say  $\{\varphi_i\}$ , we can define a (possibly infinite-dimensional) matrix  $\mathbf{A}$  associated with the bilinear form  $a(.,.)$  with the entries

$$a_{ij} = a(\varphi_j, \varphi_i).$$

If we restrict the bilinear forms to a finite dimensional subspace  $X_n$ , we denote the matrices  $\mathbf{A}$  restricted to this subspace by  $\mathbf{A}_n$ .

**4.0.10 Definition:** The two extremal eigenvalues of the matrix  $\mathbf{A}_n$  can be obtained by the maximum and minimum of the **Rayleigh quotient**

$$\Lambda(\mathbf{A}) = \max_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \quad \lambda(\mathbf{A}) = \min_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \quad (4.4)$$

The spectral condition number is

$$\kappa_n(\mathbf{A}) = \frac{\Lambda(\mathbf{A})}{\lambda(\mathbf{A})}.$$

**Remark 4.0.11.** The spectral condition number of the operator  $A$  depends on the bilinear form  $a(.,.)$  and the choice of the norm in  $X$ . On the other hand, the spectral condition number of the matrix  $\mathbf{A}$  depends on the choice of a basis of the space  $X_n$ .

## 4.1 The Richardson iteration

**4.1.1.** As a first example and prototype for all other iterative methods we consider Richardson's method, which for matrices and vectors in  $\mathbb{R}^n$  reads

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \omega_k (\mathbf{A} \mathbf{x}^{(k)} - \mathbf{b}). \quad (4.5)$$

$\omega_k$  is a relaxation parameter, which can be chosen a priori or can be changed in every step. We will for simplicity assume  $\omega_k = \omega$ .

**4.1.2 Lemma:** The error after one step of the Richardson method is given by

$$\mathbf{x}^{(k+1)} - x = \mathbf{E} \left( \mathbf{x}^{(k)} - x \right), \quad (4.6)$$

where the error propagation operator is

$$\mathbf{E} = \mathbf{I} - \omega \mathbf{A}. \quad (4.7)$$

*Proof.* Using the fact that  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ , we write

$$\mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{x}^{(k)} - \omega(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}) - \mathbf{x}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x} - \omega\mathbf{A}(\mathbf{x}^{(k)} - \mathbf{x}).$$

□

**4.1.3 Theorem:** If  $\mathbf{A}$  is symmetric, positive definite, with extremal eigenvalues  $\lambda > 0$  and  $\Lambda > 0$ , then Richardson's method converges if and only if  $0 < \omega < 2/\Lambda$ . The optimal relaxation parameter is

$$\omega_{\text{opt}} = \frac{2}{\lambda + \Lambda}, \quad (4.8)$$

which yields an optimal contraction rate of

$$\varrho_{\text{opt}} = 1 - \frac{2\lambda}{\lambda + \Lambda} = \frac{\Lambda - \lambda}{\Lambda + \lambda} = \frac{\kappa - 1}{\kappa + 1} = 1 - \frac{2}{\kappa} + \mathcal{O}(\kappa^{-2}), \quad (4.9)$$

where  $\kappa = \Lambda/\lambda$  is the so called **spectral condition number**.

*Proof.* Convergence of this method is analyzed through the Banach fixed-point theorem, which requires contraction property of the matrix  $\mathbf{M} = \mathbf{I} - \omega\mathbf{A}$ . Alternatively, we studied a theorem that states, that a matrix iteration converges if and only if the spectral radius

$$\varrho(\mathbf{M}) = \max |\lambda(\mathbf{M})| < 1,$$

the maximum absolute value of the eigenvalues of  $\mathbf{M}$  is strictly less than one.

If  $\mathbf{A}$  is symmetric, positive definite, with eigenvalues  $\lambda_i > 0$ , we have that

$$\varrho(\mathbf{M}) = \max_i |1 - \omega\lambda_i|. \quad (4.10)$$



Let the extremal eigenvalues be determined by the minimum and maximum of the Rayleigh quotient,

$$\lambda = \min_{x \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}, \quad \text{and} \quad \Lambda = \max_{x \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \quad (4.11)$$

Then, equation (4.10) yields that the method converges for  $0 < \omega < 2/\Lambda$ . Furthermore, for  $1/\Lambda \leq \omega \leq 2/\Lambda$  we have

$$\varrho(\mathbf{M}) = \max\{-1 + \omega\Lambda, 1 - \omega\lambda\}.$$

The optimal parameter  $\omega$  is the one where both values are equal and thus (4.8) and (4.9) hold.  $\square$

**4.1.4.** The analysis of finite element methods shows that it is beneficial to give up the focus on finite dimensional spaces and rather use theory that applies to separable Hilbert spaces. If results can be obtained in this context, they can easily be restricted to finite dimensional subspaces and thus become uniform with respect to the mesh parameter. Thus, we will first reformulate Richardson's method for this case and then derive convergence estimates.

**4.1.5.** Elements of an abstract Hilbert space  $X$  will be denoted by  $u, v, w$ , etc. On the other hand, coefficient vectors in  $\mathbb{R}^n$  are denoted by letters  $\mathbf{x}, \mathbf{y}, \mathbf{z}$ , etc.

**4.1.6 Definition:** Let  $X$  be a Hilbert space with inner product  $\langle \cdot, \cdot \rangle$ . Let  $a(\cdot, \cdot)$  be a second bilinear form on  $X$  and the domain of its operator is  $V$ . Then, for any right hand side  $f \in V$  and any start vector  $u^{(0)} \in V$ , **Richardson's method** is defined by the iteration

$$\langle u^{(k+1)}, v \rangle = \langle u^{(k)}, v \rangle - \omega_k (a(u^{(k)}, v) - \langle f, v \rangle), \quad \forall v \in X. \quad (4.12)$$

$\omega_k$  is a suitable relaxation parameter, chosen such that the method converges.

**4.1.7.** The scalar products in (4.12) become necessary, since different from the case in  $\mathbb{R}^n$ , the result of applying the bilinear form  $a(\cdot, \cdot)$  to  $u^{(k)}$  in the first argument yields a linear form on  $X$ . In order to convert this to a vector in  $X$ , we have to apply the isomorphism induced by the Riesz representation theorem.

**4.1.8 Theorem:** Let the bilinear form  $a(\cdot, \cdot)$  be bounded and elliptic on  $X \times X$ , namely, let there exist positive constants  $\Lambda$  and  $\lambda$  such that for all  $u, v \in X$  there holds

$$a(u, v) \leq \Lambda \|u\| \|v\|, \quad a(u, u) \geq \lambda \|u\|^2. \quad (4.13)$$

Then, Richardson's iteration converges for  $\omega_k = \omega$  for any  $\omega \in (0, 2\lambda/\Lambda^2)$ .

*Proof.* We define the iteration operator  $T$  as the solution operator of equation (4.12), namely  $Tu^{(k)} := u^{(k+1)}$ . We have to prove that  $T$  is a contraction on  $X$  under the assumptions of the theorem.

For two arbitrary vectors  $u^1, u^2 \in X$ , let  $w = u^1 - u^2$  be their difference. Due to linearity, we have  $Tw = Tu^1 - Tu^2$  and

$$\langle Tw, v \rangle = \langle w, v \rangle - \omega a(w, v) = \langle w - \omega Aw, v \rangle.$$

Using  $v = Tw$  as a test function, we obtain

$$\begin{aligned} \|Tw\|^2 &= \langle w - \omega Aw, w - \omega Aw \rangle \\ &= \|w\|^2 - 2\omega a(w, w) + \omega^2 \|Aw\|^2 \\ &\leq \|w\|^2 - 2\lambda\omega \|w\|^2 + \Lambda^2\omega^2 \|w\|^2 \\ &= \underbrace{(1 - 2\lambda\omega + \Lambda^2\omega^2)}_{=: \varrho(\omega)} \|w\|^2. \end{aligned}$$

The function  $\varrho(\omega)$  is a parabola open to the top, which at zero equals one and has a negative derivative. Thus, it is less than one for small positive values of  $\omega$ . The other point where  $\varrho(\omega) = 1$  is  $\omega = 2\lambda/\Lambda^2$ .  $\square$

**Remark 4.1.9.** The condition on  $\omega$  in Theorem 4.1.8 is more restrictive than in Theorem 4.1.3, since  $\lambda/\Lambda \leq 1$ . This is due to the fact, that in Theorem 4.1.3 we assume symmetry, and thus orthogonal diagonalizability of the matrix  $\mathbf{A}$ . With similar assumptions, Theorem 4.1.8 could be made sharper.

**Remark 4.1.10.** It is clear that the boundedness and ellipticity estimates (4.13) hold for any finite dimensional subspace  $X_n \subset X$ , and thus the convergence estimate (4.9) becomes independent of  $n$ .

More interesting and also more common is the case where the bilinear form  $a(., .)$  is unbounded on  $X$ . While it is still bounded on each finite subspace  $X_n$ , this bound cannot be independent of  $n$  if the sequence  $\{X_n\}$  approximates  $X$ .

**4.1.11 Definition:** We define an operator  $B : X \rightarrow X^*$  such that  $Bu = b(u, .) := \langle u, . \rangle$ . By the Riesz representation theorem, there is a continuous inverse operator  $B^{-1} : X^* \rightarrow X$ , which is often called **Riesz isomorphism**.

**4.1.12 Definition:** When we apply Richardson's method as in (4.12) on a computer, each step involves a multiplication with the matrix  $\mathbf{A}$ , but an inversion of the matrix  $\mathbf{B}$ , corresponding to the iteration

$$\mathbf{B}\mathbf{x}^{(k+1)} = \mathbf{B}\mathbf{x}^{(k)} - \omega_k(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}),$$

or equivalently,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \omega_k \mathbf{B}^{-1}(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{b}). \quad (4.14)$$

**Remark 4.1.13.** The iteration in (4.14) is commonly referred to as **preconditioned Richardson iteration** and  $\mathbf{B}^{-1}$  as the **preconditioner**. Note that by introducing the iteration in its weak form (4.12), the preconditioner arrives naturally and with necessity.

The goal of this chapter is finding preconditioners  $\mathbf{B}^{-1}$ , or equivalently inner products  $\langle \cdot, \cdot \rangle$ , such that the bilinear form  $a(\cdot, \cdot)$  is bounded and the condition number  $\kappa = \Lambda/\lambda$  is small.

In order to reduce (or increase) confusion, we will refer to the inner product that we search in order to bound the condition number as  $b(\cdot, \cdot)$  instead of  $\langle \cdot, \cdot \rangle$ , this way separating the Hilbert space  $X$  more clearly from the task of preconditioning. Thus, the operator  $B$  and the matrix  $\mathbf{B}$  will be associated with a bilinear form  $b(\cdot, \cdot)$  and the final version of the preconditioned Richardson iteration is

$$b(u^{(k+1)}, v) = b(u^{(k)}, v) - \omega_k(a(u^{(k)}, v) - f(v)), \quad \forall v \in X, \quad (4.15)$$

or in operator form

$$u^{(k+1)} = u^{(k)} - \omega_k B^{-1}(Au^{(k)} - f). \quad (4.16)$$

**Remark 4.1.14.** The space  $X$  and its inner product does not appear anymore in this formulation, since the bilinear form  $b(\cdot, \cdot)$  has replaced it. Thus, finding a preconditioner also amounts to changing the space in which we iterate. This is reflected by the following:

**Corollary 4.1.15.** *Let the symmetric bilinear forms  $a(\cdot, \cdot)$  and  $b(\cdot, \cdot)$  in the Richardson iteration (4.15) be both bounded and positive definite on the same space  $V$  and fulfill the **spectral equivalence** relation*

$$\lambda b(u, u) \leq a(u, u) \leq \Lambda b(u, u), \quad \forall u \in V. \quad (4.17)$$

*Then, if  $\omega_k \equiv \omega \in (0, 2\Lambda)$ , the iteration is a contraction on  $V$ . The optimal contraction number is  $\varrho$  according to equation (4.9) for  $\omega$  chosen as in (4.8).*

*Proof.* This corollary is equivalent to Theorem 4.1.8 if the inner product  $\langle \cdot, \cdot \rangle$  is replaced by the bilinear form  $b(\cdot, \cdot)$ .  $\square$

**Remark 4.1.16.** Originally, the space  $V$  was chosen as the domain of  $A$  which essentially meant  $V \subset H^2(\Omega)$ , since we required  $Av \in X$ . An additional benefit of the preconditioned version is, that now  $V \subset H^1(\Omega)$  is sufficient and at least here no regularity assumption is required.

**Notation 4.1.17.** In order to distinguish different preconditioners, we will also use the notation  $\lambda(B, A)$  and  $\Lambda(B, A)$  to refer to the constants in the norm equivalence (4.17).

**Example 4.1.18.** Let us take the example (4.1). By the Poincaré-Friedrichs inequality,  $a(., .)$  is an inner product on  $X$  and thus we can choose  $\langle ., . \rangle = a(., .)$ . In particular,  $\lambda = \Lambda = 1$  and the optimal choice is  $\omega = 1$ . Then, Richardson's iteration becomes

$$a(u^{(k+1)}, v) = a(u^{(k)}, v) - (a(u^{(k)}, v) - f(v)) = f(v), \quad \forall v \in X,$$

which converges in a single step, but we have to solve the original equation for  $u$ . Thus, either the inversion of the matrix  $A_n$  is trivial on each finite dimensional subspace  $X_n$ , or the method is useless. With usual finite element bases, the latter is true.

**Example 4.1.19.** In the other extreme, we would like to use the  $\mathbb{R}^n$  or  $L^2$  inner product on  $X_n$  or  $X$ , such that the Riesz isomorphism is easily computable. But then, the bilinear form  $a(., .)$  is unbounded on  $X$ . Thus, while for each finite  $n$ , the condition number  $\kappa_n = \Lambda_n/\lambda_n$  exists, it converges to infinity if  $n \rightarrow \infty$ .

## 4.2 The conjugate gradient method

**4.2.1.** Relying on Hilbert space structure more than Richardson's iteration is the conjugate gradient method (cg), since it uses orthogonal search directions. Nevertheless, it also relies on constructing search directions from residuals, such that a Riesz isomorphism enters the same way as before and can then be used for preconditioning.

The beauty of the conjugate gradient method is, that it is parameter and tuning free, and it converges considerably faster than a linear iteration method.

**4.2.2 Definition (Method of steepest descent):** Let  $a(.,.)$  be a symmetric, positive definite bilinear form on  $V$ . Then, the method of **steepest descent** for the energy functional

$$E(v) = \frac{1}{2}a(v, v) - f(v), \quad (4.18)$$

reads: given an initial vector  $u^{(0)}$ , compute for  $k \geq 0$  iteratively

$$\langle p^{(k)}, v \rangle = \langle -\nabla E(u^{(k)}), v \rangle_{V^* \times V} \quad \forall v \in V \quad (4.19)$$

$$\alpha_k = \operatorname{argmin}_{\alpha > 0} E(u^{(k)} + \alpha p^{(k)}) \quad (4.20)$$

$$u^{(k+1)} = u^{(k)} + \alpha_k p^{(k)} \quad (4.21)$$

**4.2.3 Lemma:** Let  $u \in V$  be the unique minimizer of the energy functional  $E(.)$ . Then, there holds for all  $k$

$$a(u^{(k+1)} - u, p^{(k)}) = a(u^{(k+1)}, p^{(k)}) - f(p^{(k)}) = 0. \quad (4.22)$$

With  $r^{(k)} = f - a(u^{(k)}, .) \in V^*$ , there holds

$$\alpha_k = \frac{\langle r^{(k)}, p^{(k)} \rangle_{V^* \times V}}{a(p^{(k)}, p^{(k)})}. \quad (4.23)$$

**4.2.4 Definition:** Let  $V$  be a Hilbert space and  $V^*$  its dual. The **conjugate gradient method** for an iteration vector  $u^{(k)} \in V$  involves the residuals  $r^{(k)} \in V^*$  as well as the update direction  $p^{(k)} \in V$  and the auxiliary vector  $w^{(k)} \in V$ . It consists of the steps

1. Initialization: for  $f$  and  $u^{(0)}$  given, compute  $r^{(0)} = f - a(u^{(0)}, \cdot)$  and

$$\langle p^{(0)}, v \rangle = \langle w^{(0)}, v \rangle = \langle r^{(0)}, v \rangle_{V^* \times V} \quad \forall v \in V.$$

2. Iteration step: for  $u^{(k)}, r^{(k)}, w^{(k)}$ , and  $p^{(k)}$  given, compute

$$\begin{aligned} u^{(k+1)} &= u^{(k)} + \alpha_k p^{(k)} & \alpha_k &= \frac{\langle r^{(k)}, w^{(k)} \rangle_{V^* \times V}}{a(p^{(k)}, p^{(k)})} \\ r^{(k+1)} &= r^{(k)} - \alpha_k a(p^{(k)}, \cdot) \\ \langle w^{(k+1)}, v \rangle &= \langle r^{(k+1)}, v \rangle_{V^* \times V} & \forall v \in V \\ p^{(k+1)} &= w^{(k+1)} + \beta_k p^{(k)} & \beta_k &= \frac{\langle r^{(k+1)}, w^{(k+1)} \rangle_{V^* \times V}}{\langle r^{(k)}, w^{(k)} \rangle_{V^* \times V}} \end{aligned}$$

**4.2.5 Definition:** The **preconditioned cg method** is obtained from above algorithm by reinterpreting the Riesz isomorphism in the computation of  $w^{(k+1)}$  as a preconditioning operation, much alike Definition 4.1.12 of the preconditioned Richardson iteration. Thus, the line defining  $w^{(k+1)}$  is replaced by

$$b(w^{(k+1)}, v) = r^{(k+1)}(v) \quad \forall v \in V.$$

Here, like there, the preconditioner enters naturally from the weak form of the algorithm.

**4.2.6 Definition:** The  $n$ th **Krylov space** as subspace of the Hilbert space  $V$  with inner product  $b(\cdot, \cdot)$  of the operator  $A$  and seed vector  $w \in V$  is

$$\mathcal{K}_n = \mathcal{K}_n(B^{-1}A, w) = \text{span} \{w, B^{-1}Aw, (B^{-1}A)^2w, \dots, (B^{-1}A)^{n-1}w\}. \quad (4.24)$$

**4.2.7 Lemma:** The vectors generated by the conjugate gradient iteration have the following properties: either  $u^{(k)}$  is the solution to the linear system or

$$r^{(k)} = f - a(u^{(k)}, \cdot) \quad (4.25)$$

$$\langle r^{(k)}, w^{(k)} \rangle_{V^* \times V} = \langle r^{(k)}, p^{(k)} \rangle_{V^* \times V} \quad (4.26)$$

$$\langle r^{(k)}, p^{(j)} \rangle = 0 \quad j < k \quad (4.27)$$

$$a(p^{(k)}, p^{(j)}) = 0 \quad j < k \quad (4.28)$$

$$\langle r^{(k)}, r^{(j)} \rangle = 0 \quad j < k \quad (4.29)$$

**4.2.8 Lemma:** The iterates of the cg method have the following minimization properties:

$$\begin{aligned} \|u^{(k)} - u\|_A &= \min_{v \in \mathcal{K}_k} \|u^{(0)} + v - u\|_A \\ &= \min_{\substack{p \in P_{n-1} \\ p(0)=1}} \|u^{(0)} + p(B^{-1}A)w - u\|_A. \end{aligned} \quad (4.30)$$

**4.2.9 Theorem:** Let the bilinear form  $a(\cdot, \cdot)$  be symmetric, and let the spectral equivalence hold. Then, the preconditioned cg method converges and we have the estimate

$$\|u^{(k)} - u\|_A \leq 2 \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|u^{(0)} - u\|_A. \quad (4.31)$$

Here,  $\kappa = \Lambda/\lambda$  is the spectral condition number of the preconditioned problem.

### 4.3 Condition numbers of finite element matrices

**4.3.1 Notation:** In asymptotic estimates, we will use the notation

$$a \lesssim b \quad \text{and} \quad b \gtrsim a,$$

if there is a constant  $c$  independent of the relevant quantities, such that

$$a \leq c b \quad \text{and} \quad c b \geq a.$$

We use

$$a \simeq b,$$

if both hold. The latter is more precise than  $a = \mathcal{O}(b)$ , since it implies the same asymptotic order.

**Example 4.3.2.** The interpolation estimate (2.50) could be written as

$$\|u - I_h u\|_{m;h} \lesssim h^{k+1-m} |u|_{k+1;h},$$

instead of

$$\|u - I_h u\|_{m;h} \leq c h^{k+1-m} |u|_{k+1;h}.$$

The statement of the equivalence of two norms becomes

$$\|\cdot\|_X \simeq \|\cdot\|_Y,$$

while

$$\|\cdot\|_X \lesssim \|\cdot\|_Y$$

states that the  $Y$ -norm is not weaker than the  $X$ -norm.

**4.3.3 Definition:** Let  $\mathbb{T}_h$  be a mesh on the domain  $\Omega$  and  $V_h$  a finite element space with basis  $\{\varphi_i\}$ . Then, the matrix associated with the  $L^2$ -inner product,

$$m_{ij} = \int_{\Omega} \varphi_j(x) \varphi_i(x) \, dx,$$

is called **mass matrix**. The matrix associated with the bilinear form of the equation is called **stiffness matrix**,

$$a_{ij} = a(\varphi_j, \varphi_i).$$



**4.3.4 Lemma:** Let  $\{\varphi_i\}$  be the basis of a finite element shape function space on a quasi-uniform family of meshes  $\{\mathbb{T}_h\}$ . Let  $\mathbf{M}_h$  be the mass matrix. Then,

$$\Lambda(\mathbf{M}) \simeq h^d \simeq \lambda(\mathbf{M})$$

Therefore, the condition number is (asymptotically in  $h$ )

$$\kappa(\mathbf{M}) \simeq 1.$$

*Proof.* For any mesh cell  $T \in \mathbb{T}_h$ , let  $\mathbf{x}_T$  be the entries of the vector  $\mathbf{x}$  which belong to node values of the cell  $T$ . Let  $\mathbf{M}_T$  be the cell mass matrix obtained by restricting the  $L^2$ -inner product to  $T$ . Then,

$$\|u_h\|_{L^2(\Omega)}^2 \mathbf{x}^T \mathbf{M}_h \mathbf{x} = \sum_{T \in \mathbb{T}_h} \mathbf{x}_T^T \mathbf{M}_T \mathbf{x}_T \begin{cases} \geq \min_{T \in \mathbb{T}_h} \frac{\mathbf{x}_T^T \mathbf{M}_T \mathbf{x}_T}{|\mathbf{x}_T|} \sum_{T \in \mathbb{T}_h} |\mathbf{x}_T|^2 \geq \lambda(\mathbf{M}_T) |\mathbf{x}|^2, \\ \leq \max_{T \in \mathbb{T}_h} \frac{\mathbf{x}_T^T \mathbf{M}_T \mathbf{x}_T}{|\mathbf{x}_T|} \sum_{T \in \mathbb{T}_h} |\mathbf{x}_T|^2 \leq c \lambda(\mathbf{M}_T) |\mathbf{x}|^2, \end{cases}$$

where the constant in the upper bound is due to degrees of freedom shared by different elements. Dividing by  $|\mathbf{x}|^2$ , we obtain

$$\lambda(\mathbf{M}_T) \leq \frac{\mathbf{x}^T \mathbf{M}_h \mathbf{x}}{|\mathbf{x}|^2} \leq c \lambda(\mathbf{M}_T). \quad (4.32)$$

In order to estimate the eigenvalues of  $\mathbf{M}_T$ , we note that for a unisolvent element, the norms  $|\mathbf{x}_T|$  and  $\|u\|_{0,T}$  are equivalent on the reference cell, and the  $L^2$ -norm scales with  $h^d$  when transforming to the real cell  $T$ . Thus, we have  $\lambda(\mathbf{M}_h) = \mathcal{O}(h^d) = \Lambda(\mathbf{M}_h)$ .  $\square$

**4.3.5 Corollary:** Let  $\mathbb{T}_h$  be a shape-regular mesh with cell sizes ranging between the minimum  $h$  and the maximum  $H$ . Then, we have

$$\begin{aligned} \Lambda(\mathbf{M}) &\simeq H^d \\ \lambda(\mathbf{M}) &\simeq h^d \\ \kappa(\mathbf{M}) &\simeq \left(\frac{H}{h}\right)^d \end{aligned}$$

**Remark 4.3.6.** We have only computed the dependence of the condition number on the mesh size  $h$ , not on the polynomial degree of the shape function space. Indeed, the used equivalence between the  $L^2$ -norm of a polynomial on the reference cell and the Euclidean norm of its coefficient vector depends not only on the degree of the shape function space, but on its basis.

**4.3.7 Theorem:** Let  $\{\varphi_i\}$  be the nodal basis of a finite element shape function space on a quasi-uniform family of meshes  $\{\mathbb{T}_h\}$ . Let  $\mathbf{A}_h$  be the stiffness matrix of the Laplacian with Dirichlet boundary values. Then,

$$\Lambda(\mathbf{A}) \simeq h^{d-2}, \quad \lambda(\mathbf{A}) \simeq h^d$$

Therefore, the condition number is

$$\kappa(\mathbf{M}) \simeq h^{-2}.$$

*Proof.* First, we investigate the largest eigenvalue. By the scaling argument, there holds  $a_{ij} \simeq h^{d-2}$ . Since by the Gershgorin theorem all eigenvalues are bounded by the sum of the elements in a row, and this number is bounded uniformly on shape regular meshes, we have  $\Lambda(\mathbf{A}_h) \lesssim h^{d-2}$ . Now, we choose a particular example, namely a vector  $\mathbf{x}$  with a single one and all other entries zero. The corresponding finite element function  $u_h$  has maximum one and support of radius  $h$ . Therefore,

$$\mathbf{x}^T \mathbf{A}_h \mathbf{x} = \int_{\Omega} |\nabla u_h|^2 d\mathbf{x} \simeq h^{d-2}.$$

Thus,  $\Lambda(\mathbf{A}_h) \gtrsim h^{d-2}$ . In order to estimate the smallest eigenvalue, we use the Rayleigh quotient and the finite element function  $u_h$  corresponding to the coefficient vector  $\mathbf{x}$  to obtain

$$\begin{aligned} \lambda(\mathbf{A}_h) &= \min_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A}_h \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \min_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{A}_h \mathbf{x}}{(u_h, u_h)_{L^2}} \frac{(u_h, u_h)_L^2}{\mathbf{x}^T \mathbf{x}} \\ &= \min_{\mathbf{x} \in \mathbb{R}^n} \left( \frac{a(u_h, u_h)}{(u_h, u_h)} \frac{\mathbf{x}^T \mathbf{M}_h \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \right) \geq \min_{u_h \in V_h} \frac{a(u_h, u_h)}{(u_h, u_h)_L^2} \min_{\mathbf{x} \in \mathbb{R}^n} \frac{\mathbf{x}^T \mathbf{M}_h \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \end{aligned} \quad (4.33)$$

Thus, the smallest eigenvalue of  $\mathbf{A}_h$  is bounded by the product of the smallest eigenvalue of the mass matrix and a value  $\mu$  given by

$$\min_{u_h \in V_h} \frac{a(u_h, u_h)}{(u_h, u_h)_L^2} \geq \min_{u \in V} \frac{a(u, u)}{(u, u)_L^2} =: \mu. \quad (4.34)$$

Here, we need a result from the spectral analysis of the Laplacian. When we define the variational eigenvalue problem

$$a(u, v) = \lambda(u, v) \quad \forall v \in V = H_0^1(\Omega), \quad (4.35)$$

then, there is a positive minimal solution  $\mu$ , the first eigenvalue of the Laplacian, and an associated eigenfunction  $u$ , such that the equation holds for all  $v$ . Therefore, the eigenvalue is bounded with  $\lambda(\mathbf{A}_h) \gtrsim h^{-d}$ . Furthermore, it can be

shown that this function is smooth and does not oscillate. This actually follows from the regularity result in Theorem 1.4.6. Therefore, the solution  $u_h \in V_h$  to

$$a(u_h, v_h) = a(u, v_h) \quad \forall v_h \in V_h, \quad (4.36)$$

converges to  $u$  in  $H^1$  and  $L^2$ . With our standard estimates, we obtain

$$(u, u_h) = (u_h, u_h) + (u - u_h, u_h) = (u_h, u_h) + r_h(u), \quad (4.37)$$

where  $r_h(u) \rightarrow 0$  as  $h \rightarrow 0$ . Hence,

$$a(u_h, u_h) = a(u, u_h) = \mu(u, u_h) \rightarrow \mu(u_h, u_h). \quad (4.38)$$

We conclude that the lower bound for the eigenvalue is sharp and the theorem holds.  $\square$

## 4.4 Multigrid methods

**Example 4.4.1.** Smoothing property of the Richardson method with  $\omega = 1/2$  for the one-dimensional Laplacian.

### 4.4.2 Algorithm (The Multigrid iteration $\text{MGM}(\ell, u_\ell^k, b_\ell)$ ):

1. Pre-smoothing: let  $u_\ell^{k,0} = u_\ell^k$  and for  $i = 0, \dots, \nu_1 - 1$  do

$$u_\ell^{k,i+1} = u_\ell^{k,i} + \omega(b_\ell - A_\ell u_\ell^{k,i}). \quad (4.39)$$

2. Coarse grid correction: compute  $b_{\ell-1} = r(b_\ell - A_\ell u_\ell^{k,\nu_1})$  and

- (a) If  $\ell = 1$  solve  $A_0 u_0^\mu = b_0$  exactly

- (b) If  $\ell > 1$ , let  $u_{\ell-1}^0 = 0$  and for  $i = 0, \dots, \mu - 1$  do

$$u_{\ell-1}^{i+1} = \text{MGM}(\ell - 1, u_{\ell-1}^i, b_{\ell-1}).$$

Then, add  $u_\ell^{k,\nu_1+1} = u_\ell^{k,\nu_1} + p u_{\ell-1}^\mu$ .

3. Post-smoothing: for  $i = \nu_1 + 1, \dots, \nu_1 + \nu_2$  do

$$u_\ell^{k,i+1} = u_\ell^{k,i} + \omega(b_\ell - A_\ell u_\ell^{k,i}). \quad (4.40)$$

and let  $u_\ell^{k+1} = u_\ell^{k,\nu_1+\nu_2+1}$ .

**Remark 4.4.3.** The multigrid method for  $\mu = 1$  is called the V-cycle, for  $\mu = 2$  it is the W-cycle. The generalization where  $\mu$  depends on  $\ell$ , typically in the form  $2^{L-\ell}$ , where  $L$  is the finest level, is called the variable V-cycle.

**Remark 4.4.4.** The Richardson method in pre- and post-smoothing can be replaced by other relaxation methods like Jacobi, Gauss-Seidel, SOR, or SSOR. Even more complex operations may be considered.

If the smoother, like SOR, is not symmetric, it is possible to apply the transpose (backward SOR instead of forward) in post-smoothing to obtain a method which is symmetric.

**4.4.5 Lemma:** Let  $S_\ell$  be the error propagation operator of the smoother. Then, the error propagation operator of a single step of the multigrid iteration is defined recursively by

$$I - M_\ell = S_\ell^{\nu_2} (I - pM_{\ell-1}^\mu rA) S_\ell^{\nu_1}, \quad (4.41)$$

and  $M_0 = A^{-1}$ .

*Proof.* We prove for  $k = 0$ , which proves for every step and omit the superscript  $k$ . First, observe that for the exact solution  $u_\ell$  of  $A_\ell u_\ell = b_\ell$  there holds

$$u_\ell - u_\ell^{i+1} = S(u_\ell - u_\ell^{i+1}) \quad i = i = 0, \dots, \nu_1 - 1, \nu_1 + 1, \dots, \nu_1 + \nu_2.$$

Thus,

$$u_\ell - u_\ell^{\nu_1} = S^{\nu_1}(u_\ell - u_\ell^0), \quad u_\ell - u_\ell^{\nu_1 + \nu_2 + 1} = S^{\nu_2}(u_\ell - u_\ell^{\nu_1 + 1}),$$

We continue by induction. For  $\ell = 1$ , we have

$$u_1 - u_1^{\nu_1 + 1} = u_1 - u_1^{\nu_1} - pu_0^\mu = u_1 - u_1^{\nu_1} - pA_0^{-1}rA_1(u_1 - u_1^{\nu_1}).$$

Combining with the error propagation of pre and post smoothing, we get

$$u_1 - u_1^{\nu_1 + \nu_2 + 1} = S_\ell^{\nu_2} (I - pA_0^{-1}rA_1) S_\ell^{\nu_1} (u_1 - u_1^0).$$

□

## Chapter 5

# Discontinuous Galerkin methods

### 5.1 Nitsche's method

5.1.1. Let us consider the inhomogeneous Dirichlet boundary value problem

$$\begin{aligned} -\Delta u &= f & \text{in } \Omega \\ u &= u^D & \text{on } \partial\Omega. \end{aligned} \quad (5.1)$$

We already know that for  $u^D \equiv 0$  we can discretize this problem by choosing a finite element space  $V_h \subset H_0^1(\Omega)$ . For general  $u^D$ , there are two options:

1. Compute an arbitrary “lifting”  $u^D \in H^1(\Omega)$  such that it is equal to  $u^D$  on the boundary and compute  $w = u - u^D \in H_0^1(\Omega)$  as the solution to the weak formulation

$$\int_{\Omega} \nabla w \cdot \nabla v \, dx = \int_{\Omega} f v \, dx + \int_{\Omega} \nabla u^D \cdot \nabla v \, dx. \quad (5.2)$$

2. Compute an interpolation or projection  $u_h^D$  of the boundary data  $u^D$ . Then, eliminate each row of the matrix corresponding to a degree of freedom on the boundary. In particular, let  $i$  be the index of such a degree of freedom and let  $k$  be an index corresponding to an interior degree of freedom not constraint by a boundary condition, but such that  $a_{ik} \neq 0 \neq a_{ki}$ . Then, replace the rows

$$\begin{aligned} \cdots + a_{ii}u_i + \cdots + a_{ik}u_k + \cdots &= f_i \\ \cdots + a_{ki}u_i + \cdots + a_{kk}u_k + \cdots &= f_k \end{aligned} \quad (5.3)$$

by the rows

$$\begin{aligned} u_i &= u_i^D \\ \cdots + 0 + \cdots + a_{kk}u_k + \cdots &= f_k - a_{ki}u_i^D \end{aligned} \quad (5.4)$$

The first option introduces the complication of finding a function in  $H^1(\Omega)$ , which cannot be achieved automatically. The second can be implemented in an automatic way, but it complicates code, in particular for nonlinear problems.

A completely different approach modifying the bilinear form was first suggested in the 60s and then perfected by Joachim Nitsche in 1971. In this section, we motivate this method and derive its error estimates. Its key feature is the transition from  $V_h \subset H_0^1(\Omega)$  to  $V_h \subset H^1(\Omega)$ .

**5.1.2.** If we simply derive our weak formulation in  $H^1(\Omega)$ , we end up with an additional boundary term

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = - \int_{\Omega} \Delta u v \, dx + \int_{\partial\Omega} \partial_n u v \, ds. \quad (5.5)$$

Thus, we obtain the natural boundary condition  $\partial_n u = 0$ , which is not consistent with the original BVP. The first step for deriving Nitsche's method is subtracting this boundary term on both sides. The result is the equation

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \partial_n u v \, ds = \int_{\Omega} f v \, dx \quad \forall v \in H^1(\Omega). \quad (5.6)$$

We observe that the left hand side vanishes for any constant function  $u$ . Thus, we do not have unique solvability and we will have to fix this problem. Furthermore, the boundary data  $u^D$  does not appear in this formulation. We enforce  $u = u^D$  in our formulation by a so called “penalty term” with penalty parameter  $\alpha$ , modifying (5.6) to

$$\begin{aligned} \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \partial_n u v \, ds + \int_{\partial\Omega} \alpha u v \, ds \\ = \int_{\Omega} f v \, dx + \int_{\partial\Omega} \alpha u^D v \, ds \quad \forall v \in H^1(\Omega). \end{aligned} \quad (5.7)$$

Integrating by parts, we see that  $u$  is a solution to this weak formulation. Following Nitsche, we make one additional modification which restores the symmetry of our form. We obtain the weak formulation

$$a_h(u, v) = f_h(v) \quad \forall v \in H^1(\Omega), \quad (5.8)$$

where

$$\begin{aligned} a_h(u, v) &= \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \partial_n u v \, ds - \int_{\partial\Omega} \partial_n v u \, ds + \int_{\partial\Omega} \alpha u v \, ds, \\ f_h(v) &= \int_{\Omega} f v \, dx - \int_{\partial\Omega} \partial_n v u^D \, ds + \int_{\partial\Omega} \alpha u^D v \, ds. \end{aligned} \quad (5.9)$$

We abbreviate this equation to

**Remark 5.1.3.** Unfortunately, the problem (5.9) is not well-posed for any finite parameter  $\alpha$ . Thus, it cannot be used to determine  $u \in H^1(\Omega)$ . Nevertheless, we can establish well-posedness on discrete spaces  $V_h$  in order to compute a discrete solution  $u_h$  and use the fact that  $u$  is already determined by the continuous problem. Our immediate goals are thus:

1. Establish the assumptions of the Lax-Milgram theorem on  $V_h$ , which in this case involves a suitable new norm for measuring the error.
2. Establish a relation between the discrete and continuous solution replacing Galerkin orthogonality.
3. Deriving error estimates in suitable norms.

**Notation 5.1.4.** From now on, we will use the inner product notation

$$(u, v) \equiv \int_{\Omega} uv \, dx \quad (\nabla u, \nabla v) \equiv \int_{\Omega} \nabla u \cdot \nabla v \, dx,$$

on  $\Omega$  as well as

$$\langle u, v \rangle \equiv \int_{\partial\Omega} uv \, ds,$$

on its boundary.

**Definition 5.1.5.** A discrete problem (5.9) is called **consistent**, if for the solution  $u$  of the BVP (5.1) there holds

$$a_h(u, v_h) = f_h(v_h) \quad \forall v_h \in V_h. \quad (5.10)$$

**Corollary 5.1.6.** Let  $u \in H^1(\Omega)$  be the weak solution to (5.1) in the sense of (5.2). Then, the discrete problem (5.9) is consistent.

*Proof.* Since  $u \in H^1(\Omega)$  and  $v_h \in V_h$ , all boundary terms in  $a_h(u, v - h)$  and  $f_h(v_h)$  are well-defined and consistency follows from  $u = u^D$  in the sense of  $L^2(\partial\Omega)$ .  $\square$

**Definition 5.1.7.** The problem dependent norm used for the analysis of Nitsche's method is defined by

$$\|v\|^2 = (\nabla v, \nabla v) + \langle \alpha v, v \rangle. \quad (5.11)$$

This choice is justified by

**Lemma 5.1.8.** Let  $V_h \subset V = H^1(\Omega)$  be a piecewise polynomial finite element space on a shape-regular mesh  $\mathbb{T}_h$ . Then, if  $\alpha$  sufficiently large, there exist constants  $M$  and  $\gamma$  such that

$$\begin{aligned} a_h(u_h, v_h) &\leq M \|u_h\| \|v_h\| \\ a_h(u_h, u_h) &\geq \gamma \|u_h\|^2. \end{aligned}$$

*Proof.* Key for the proof is the inverse trace estimate

$$|v|_{H^1(\partial T)} \leq ch^{-1/2}|v|_{H^1(T)},$$

which holds with a constant  $c$  depending on shape regularity and polynomial degree. Thus, for a cell  $T$  at the boundary, there holds

$$\begin{aligned} |\langle \partial_n u_h, v_h \rangle_{\partial T \cap \partial \Omega}| &\leq ch^{-1/2}|u_h|_{H^1(T)} \|v_h\|_{L^2(\partial T \cap \partial \Omega)} \\ &\leq \frac{1}{4}|u_h|_{H^1(T)}^2 + \frac{c^2}{h_T} \|v_h\|_{L^2(\partial T \cap \partial \Omega)}^2, \end{aligned}$$

We apply this to the lower bound to obtain

$$a_h(u_h, u_h) \geq \left(1 - \frac{1}{2}\right) |u_h|_{H^1(\Omega)}^2 + \left(\alpha - \frac{2c^2}{h_T}\right) \|u_h\|_{L^2(\partial \Omega)}^2.$$

Choosing

$$\alpha(x) = \frac{\alpha_0}{h(x)} \geq \frac{4c^2}{h(x)}, \quad (5.12)$$

where  $h(x)$  is the size of the cell such that  $x \in \partial T$ , we obtain

$$a_h(u_h, u_h) \geq \frac{1}{2} \|u_h\|^2.$$

The proof of the upper bound follows the same fashion.  $\square$

**Corollary 5.1.9.** *Let  $\alpha$  be chosen according to equation (5.12). Then, the discrete problem (5.8) has a unique solution  $u_h \in V_h$ .*

*Proof.* According to the previous lemma, the lemma of Lax-Milgram applies to the bilinear form  $a_h(., .)$ . For the right hand side  $f_h(., .)$ , we have again because of trace estimates in  $H^1(\Omega)$  and inverse estimates in  $V_h$

$$f_h(v) \leq c (\|f\|_{0,\Omega} + \|u^D\|_{H^1(\Omega)}) \|v\|.$$

$\square$

**5.1.10 Theorem:** Let  $u \in H^{k+1}(\Omega)$  with  $k \geq 1$  be the solution of (5.1) and let  $u_h \in V_h$  be the solution to (5.8) and let the assumptions of Lemma 5.1.8 hold. Let furthermore  $\{\mathbb{T}_h\}$  be a family of quasi-uniform, shape-regular meshes of maximal cell diameter  $h$ , and let the shape function spaces contain the polynomial space  $P_k$ . Then, there holds

$$\|u - u_h\| \leq ch^k |u|_{k+1, \mathbb{T}_h}. \quad (5.13)$$



*Proof.* We begin with the triangle inequality

$$\|u - u_h\| \leq \|u - I_h u\| + \|I_h u - u_h\|.$$

The interpolation error can be estimated by

$$\begin{aligned} |u - I_h u|_{1,\Omega} &\leq h^k |u|_{2,\Omega}, \\ |u - I_h u|_{0,\partial\Omega} &\leq h^{3/2} |u|_{2,\Omega}, \\ |u - I_h u|_{1,\partial\Omega} &\leq h^{1/2} |u|_{2,\Omega}. \end{aligned} \quad (5.14)$$

The second and third estimate actually require some deeper arguments from functional analysis, which is beyond the scope of this class. It involves an intuitive notion of Sobolev spaces with non-integer derivatives. Allowing such spaces, the trace estimate becomes

$$\|u\|_{1/2,\partial\Omega} \leq c \|u\|_{1,\Omega}. \quad (5.15)$$

For the remaining error term, we use  $V_h$ -ellipticity of the discrete form and consistency to obtain

$$\gamma \|I_h u - u_h\|^2 \leq a_h(I_h u - u_h, I_h u - u_h) = a_h(I_h u - u, I_h u - u_h). \quad (5.16)$$

Using Young's inequality, we estimate the right hand side with  $\varepsilon_h = u - I_h u$  and  $\eta_h = I_h u - u_h$  on each boundary cell  $T$  with boundary edge  $E$  by

$$\begin{aligned} |(\nabla \varepsilon_h, \nabla \eta_h)_T| &\leq \frac{1}{\gamma} |\varepsilon_h|_{1,T}^2 + \frac{\gamma}{4} |\eta_h|_{1,T}^2, \\ |\langle \alpha \varepsilon_h, \eta_h \rangle_E| &\leq \frac{2}{\gamma} \|\sqrt{\alpha} \varepsilon_h\|_{0,E}^2 + \frac{\gamma}{8} \|\sqrt{\alpha} \eta_h\|_{0,E}^2, \\ |\langle \varepsilon_h, \partial_n \eta_h \rangle_E| &\leq \frac{1}{\gamma} \|\sqrt{\alpha} \varepsilon_h\|_{0,E}^2 + \frac{\gamma}{4} |\eta_h|_{1,T}^2 + \frac{\gamma}{4} \|\sqrt{\frac{\alpha_0}{h_T}} \eta_h\|_{0,E}^2, \\ |\langle \partial_n \varepsilon_h, \eta_h \rangle_E| &\leq \frac{2h_T}{\gamma \alpha_0} |\varepsilon_h|_{0,E}^2 + \frac{\gamma \alpha_0}{8h_T} \|\eta_h\|_{0,E}^2. \end{aligned} \quad (5.17)$$

Adding these over all cells, we obtain

$$\begin{aligned} &\gamma \|I_h u - u_h\|^2 \\ &\leq \frac{\gamma}{2} \|I_h u - u_h\|^2 \left( \frac{1}{\gamma} |u - I_h u|_{1,\Omega}^2 + \frac{3}{\gamma} \|\sqrt{\alpha}(u - I_h u)\|_{0,E}^2 + \frac{2h_T}{\gamma \alpha_0} |u - I_h u|_{0,E}^2 \right), \end{aligned}$$

and thus, by the interpolation estimate (5.14)

$$\|I_h u - u_h\|^2 \leq \frac{2}{\gamma} c h^{2k} |u|_{k+1,\Omega}^2.$$

□

**5.1.11 Theorem:** Assume in addition to the assumptions of Theorem 5.1.10 that the adjoint problem

$$\begin{aligned} -\Delta z &= u - u_h & \text{in } \Omega \\ z &= 0 & \text{on } \partial\Omega. \end{aligned} \quad (5.18)$$

admits elliptic regularity, namely

$$|z|_2 \leq c \|u - u_h\|_0. \quad (5.19)$$

Then, the solutions  $u$  and  $u_h$  admit the estimate

$$\|u - u_h\|_0 \leq ch^{k+1} |u|_{k+1}. \quad (5.20)$$

*Proof.* Due to symmetry of the discrete bilinear form, we have adjoint consistency:

$$a_h(v_h, z) = (u - u_h, v_h) \quad \forall v_h \in V_h. \quad (5.21)$$

From here, we proceed like in the continuous case:

$$\|u - u_h\|_0^2 = a_h(u - u_h, z) = a_h(u - u_h, z - I_h z).$$

Using the same derivation as in the previous theorem, we obtain the result.  $\square$

**Remark 5.1.12.** We chose the norm defined in (5.11) for our energy norm analysis in Theorem 5.1.10. We could have done the same using the operator norm  $\sqrt{a_h(v, v)}$ . In fact, Lemma 5.1.8 states that both norms are equivalent. Therefore, we chose the one involving less terms and providing for simpler interpolation estimates.

The “triple norm” notation  $\|\cdot\|$  is very common as a notation for problem adjusted norms.

## 5.2 The interior penalty method

**5.2.1.** We review the basic definitions necessary to describe discontinuous Galerkin (DG) methods. In particular, we need the sets of faces  $\mathbb{F}_h$  of a mesh, discontinuous piecewise polynomial spaces and broken integrals.

**5.2.2 Definition:** Let  $\mathbb{T}_h$  be a mesh of  $\Omega \subset \mathbb{R}^d$  consisting of mesh cells  $T_i$ . For every boundary facet  $F \subset \partial T_i$ , we assume<sup>a</sup> that either  $F \subset \partial\Omega$  or  $F$  is a boundary facet of another cell  $T_j$ . In the second case, we indicate this relation by labeling this facet  $F_{ij}$ . The set of all facets  $F_{ij}$  is the set of interior faces  $\mathbb{F}_h^i$ . The set of facets on the boundary is  $\mathbb{F}_h^\partial$ .

<sup>a</sup>This assumption can indeed be relaxed

**5.2.3 Definition:** The discontinuous finite element space on  $\mathbb{T}_h$  is constructed by concatenation of all shape function spaces  $P_T$  for  $T \in \mathbb{T}_h$  without additional continuity requirements:

$$V_h = \{v \in L^2(\Omega) \mid v|_T \in P_T \ \forall T \in \mathbb{T}_h\}. \quad (5.22)$$

**5.2.4 Definition:** For any set of cells  $\mathbb{T}_h$  or faces  $\mathbb{F}_h$ , we define the bilinear forms

$$(u, v)_{\mathbb{T}_h} = \sum_{T \in \mathbb{T}_h} (u, v)_T, \quad (5.23)$$

$$\langle u, v \rangle_{\mathbb{F}_h} = \sum_{F \in \mathbb{F}_h} \langle u, v \rangle_F. \quad (5.24)$$

$$(5.25)$$

**5.2.5.** We start out with the equation

$$-\Delta u = f.$$

Integrating by parts on each mesh cell yields

$$(-\Delta u, v)_T = (\nabla u, \nabla v)_T - \langle \partial_n u, v \rangle_{\partial T} = (f, v)_T.$$

We realize that the choice of discontinuous finite element spaces introduces a consistency term on the interfaces between cells and on the boundary.

On interior faces, there is the issue that  $u$  and  $\partial_n u$  actually have two values on the interface, one from the left cell and one from the right. Therefore, we have to consolidate these two values into one. To this end, we introduce the concept of a numerical flux, which constructs a single value out of these two. Thus, we introduce on the interface  $F$  between two cells  $T^+$  and  $T^-$

$$\mathcal{F}(\nabla u) = \frac{\nabla u^+ + \nabla u^-}{2} =: \{\!\!\{ \nabla u \}\!\!\}.$$

Using  $\langle \partial_n u, v \rangle = \langle \nabla u, v \mathbf{n} \rangle$  we change our point of view and instead of integrating over the boundary  $\partial T$ , we integrate over a face  $F$  between two cells  $T^+$  and  $T^-$ . Adding up integrals from both sides, we obtain the term

$$-\langle \{\!\!\{ \nabla u \}\!\!\}, v^+ \mathbf{n}^+ + v^- \mathbf{n}^- \rangle_F = -2\langle \{\!\!\{ \nabla u \}\!\!\}, \{\!\!\{ v \mathbf{n} \}\!\!\} \rangle_F.$$

On boundary faces, we simply get

$$\langle \partial_{\mathbf{n}} u, v \rangle_F.$$

Adding over all cells and faces, we obtain the equation

$$(\nabla u, \nabla v)_{\mathbb{T}_h} - 2\langle \{\!\!\{ \nabla u \}\!\!\}, \{\!\!\{ v \mathbf{n} \}\!\!\} \rangle_{\mathbb{F}_h^i} - \langle \partial_{\mathbf{n}} u, v \rangle_{\mathbb{F}_h^\partial} = (f, v)_\Omega.$$

Following the idea of Nitsche, we symmetrize this term to obtain

$$\begin{aligned} & (\nabla u, \nabla v)_{\mathbb{T}_h} - 2\langle \{\!\!\{ \nabla u \}\!\!\}, \{\!\!\{ v \mathbf{n} \}\!\!\} \rangle_{\mathbb{F}_h^i} - 2\langle \{\!\!\{ u \mathbf{n} \}\!\!\}, \{\!\!\{ \nabla v \}\!\!\} \rangle_{\mathbb{F}_h^i} \\ & - \langle \partial_{\mathbf{n}} u, v \rangle_{\mathbb{F}_h^\partial} - \langle u, \partial_{\mathbf{n}} v \rangle_{\mathbb{F}_h^\partial} = (f, v)_\Omega - \langle u^o, \partial_n v \rangle_{\mathbb{F}_h^\partial}. \end{aligned}$$

Here the second term on the right was introduced for consistency. Finally, it turns out that this method is not stable and needs stabilization by a jump term. This will be done in Definition 5.2.8. Before, we introduce the notation for averaging and jump operators.

**5.2.6 Notation:** Let  $F$  be a face between the cells  $T^+$  and  $T^-$ . Let  $\mathbf{n}^+$  and  $\mathbf{n}^- = -\mathbf{n}^+$  be the outer normal vectors of the cells at a point  $x \in F$ . For a function  $u \in V_h$ , the traces  $u^+$  and  $u_-$  of  $u$  on  $F$  taken from the cell  $T^+$  and  $T^-$  are defined as:

$$\begin{aligned} u^+(x) &= \lim_{\varepsilon \searrow 0} u(x - \varepsilon \mathbf{n}^+), \\ u^-(x) &= \lim_{\varepsilon \searrow 0} u(x - \varepsilon \mathbf{n}^-). \end{aligned}$$

We define the **averaging operator**  $\{\!\!\{ \cdot \}\!\!\}$  and the **jump operator**  $\llbracket \cdot \rrbracket$  as

$$\{\!\!\{ u \}\!\!\} = \frac{u^+ + u^-}{2}, \quad \llbracket u \rrbracket = u^+ - u^-. \quad (5.26)$$

Not that the sign of the jump of  $u$  depends on the choice of the cells  $T^+$  and  $T^-$ . It will only be used in quadratic terms.

**Remark 5.2.7.** The jump can be denoted as the mean value of the product of a function and the normal vector,

$$\llbracket u \rrbracket = 2\{\!\!\{ u \mathbf{n} \}\!\!\} \cdot \mathbf{n}^+ = -2\{\!\!\{ u \mathbf{n} \}\!\!\} \cdot \mathbf{n}^-. \quad (5.27)$$

**5.2.8 Definition:** The **interior penalty method**<sup>a</sup> uses the bilinear form

$$\begin{aligned} a_h(u, v) = & (\nabla u, \nabla v)_{\mathbb{T}_h} + \langle \sigma_h \llbracket u \rrbracket, \llbracket v \rrbracket \rangle_{\mathbb{F}_h^i} + \langle \sigma_h u, v \rangle_{\mathbb{F}_h^\partial} \\ & - 2 \langle \llbracket \nabla u \rrbracket, \llbracket v \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} - 2 \langle \llbracket u \mathbf{n} \rrbracket, \llbracket \nabla v \rrbracket \rangle_{\mathbb{F}_h^i} \\ & - \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial} - \langle u, \partial_n v \rangle_{\mathbb{F}_h^\partial}, \end{aligned} \quad (5.28)$$

and the linear form

$$f_h(v) = (f, v)_\Omega - \langle u^D, \partial_n v \rangle_{\mathbb{F}_h^\partial} + \langle \sigma_h u^D, v \rangle_{\mathbb{F}_h^\partial}, \quad (5.29)$$

where  $f$  is the right hand side of the equation and  $u^D$  the Dirichlet boundary value.

---

<sup>a</sup>Also known as symmetric interior penalty (SIPG) or IP-DG.

**5.2.9 Definition:** On the space  $V_h$  we define the norm  $\|\cdot\|_{1,h}$  by

$$\|v\|_{1,h}^2 = \sum_{T \in \mathbb{T}_h} \|\nabla v\|_T^2 + \sum_{F \in \mathbb{F}_h^i} \|\sqrt{\sigma_h} \llbracket v \rrbracket\|_F^2 + \sum_{F \in \mathbb{F}_h^\partial} \|\sqrt{\sigma_h} v\|_F^2. \quad (5.30)$$

**5.2.10 Problem:** Prove that the norm defined in (5.30) is indeed a norm on  $V_h$ .

**5.2.11 Lemma:** Let  $\mathbb{T}_h$  be shape-regular and chosen on each face  $F$  as  $\sigma_h = \sigma_0/h_F$ , where  $h_T$  is the minimal diameter of a cell adjacent to  $F$ . Then, there is a  $\sigma_0 > 0$  such that there exists a constant  $\gamma > 0$ , such that independent of  $h$  there holds

$$a_h(u_h, u_h) \geq \gamma \|u_h\|_{1,h}^2 \quad \forall u_h \in V_h. \quad (5.31)$$

**5.2.12 Problem:** Prove Lemma 5.2.11.

**5.2.13 Lemma:** Let  $f \in L^2(\Omega)$  and let the boundary conditions admit that for the solution to

$$\begin{aligned} -\Delta u &= f & \text{in } \Omega, \\ u &= u^D & \text{on } \partial\Omega, \end{aligned}$$

there holds  $u \in H^{1+\varepsilon}(\Omega)$  for a positive  $\varepsilon$ . Then, the interior penalty method is consistent, that is,

$$a_h(u, v_h) = f_h(v_h) \quad \forall v_h \in V_h. \quad (5.32)$$

*Proof.* From  $f \in L^2(\Omega)$  we deduce that  $\nabla u \in H^{\text{div}}(\Omega)$ . Thus, with the extra regularity, the traces of  $\partial_n u$  on faces are well-defined and coincide from both sides. The remainder is integration by parts.  $\square$

**5.2.14 Theorem:** For  $k \geq 1$  let  $\mathbb{P}_k \subset P_T$  and  $u \in H^{s+1}(\Omega)$  with  $1/2 \leq s \leq k$ . Then, the interior penalty method admits the error estimate

$$\|u - u_h\|_{1,h} \leq ch^s |u|_{s+1}. \quad (5.33)$$

If furthermore the boundary condition admits elliptic regularity, there holds

$$\|u - u_h\|_0 \leq ch^{s+1} |u|_{s+1}. \quad (5.34)$$

### 5.2.1 Bounded formulation in $H^1$

**5.2.15.** The interior penalty method introduced so far is  $V_h$ -elliptic and consistent, but it is not bounded on  $H^1(\Omega)$ . This was a reason, why we could not use standard techniques for the proof of the convergence result and after applying consistency had to estimate each term separately.

In this section, we will introduce a reformulation of the interior penalty method, which is equivalent to the original method on  $V_h$ , but is also bounded in  $H^1(\Omega)$ . As an unpleasant side effect, it turns out that this method is inconsistent, and we have to estimate the consistency error.

The main technique applied here is the use of lifting operators, such that the traces of derivatives on faces can be replaced by volume terms. Note that the lifting operators, while very useful for the analysis of the method, are not actually used in the implementation of the interior penalty method.

**5.2.16 Definition:** Define the auxiliary space

$$\Sigma_h = \{\tau \in L^2(\Omega; \mathbb{R}^d) \mid \forall T \in \mathbb{T}_h : \tau|_T \in \Sigma_T\}, \quad (5.35)$$

where  $\Sigma_T$  is a (possibly mapped) polynomial space chosen such that  $\nabla V_T \subset \Sigma_T$ . Then, we define the **lifting operator**

$$\mathcal{L}: V + V_h \rightarrow \Sigma_h \quad (5.36)$$

by

$$(\mathcal{L}v, \tau)_{\mathbb{T}_h} = 2\langle \{\tau\}, \{v\mathbf{n}\} \rangle_{\mathbb{F}_h^i} + \langle \tau \cdot \mathbf{n}, v \rangle_{\mathbb{F}_h^\partial}. \quad (5.37)$$

**5.2.17 Lemma:** The lifting operator is a bounded operator from  $L^2(\mathbb{F}_h)$  to  $\Sigma_h$ , such that

$$\|\mathcal{L}v\|_{L^2(\Omega)} \leq c \left\| \frac{1}{\sqrt{h}} \llbracket v \rrbracket \right\|_{\mathbb{F}_h^i} + \left\| \frac{1}{\sqrt{h}} v \right\|_{\mathbb{F}_h^\partial}. \quad (5.38)$$

In particular, it is bounded on  $H^1(\Omega)$ .

*Proof.* It is clear, that the operator is bounded on  $L^2(\mathbb{F}_h)$ , since its definition involves face integrals weighted with polynomial functions. The dependence on the mesh size is due to the standard scaling argument.  $\square$

**5.2.18 Definition:** The **interior penalty method** with lifting operators uses the bilinear form

$$\begin{aligned} a_h(u, v) = & (\nabla u, \nabla v)_{\mathbb{T}_h} - (\mathcal{L}u, \nabla v)_{\mathbb{T}_h} - (\nabla u, \mathcal{L}v)_{\mathbb{T}_h} \\ & + \langle \sigma_h \llbracket u \rrbracket, \llbracket v \rrbracket \rangle_{\mathbb{F}_h^-} + \langle \sigma_h u, v \rangle_{\mathbb{F}_h^\partial}. \end{aligned} \quad (5.39)$$

and the linear form (5.29) of the original interior penalty method. Its residual operator is

$$\text{Res}(u, v) = a_h(u, v) - (f, v). \quad (5.40)$$

**5.2.19 Lemma:** The interior penalty method in flux form (Definition 5.2.8) and in lifting form (Definition 5.2.8) coincide on the discrete space  $V_h$  if  $\Sigma_h$  is chosen such that  $\nabla V_h \subset \Sigma_h$ .

*Proof.* Since  $\nabla V_h \subset \Sigma_h$ ,  $\nabla u_h$  and  $\nabla v_h$  are valid test functions in the definition (5.37) of the lifting operator, and the equality

$$(\mathcal{L}u_h, \nabla v_h)_{\mathbb{T}_h} = 2\langle \llbracket u_h \mathbf{n} \rrbracket, \llbracket \nabla v_h \rrbracket \rangle_{\mathbb{F}_h^i} + \langle u_h, \partial_n v_h \rangle_{\mathbb{F}_h^\partial}.$$

□

**5.2.20 Definition:** Let  $V \subset H^1(\Omega)$  and let  $u, u^* \in V$  solve the primal and dual problems

$$a(u, v) = f(v), \quad a(v, u^*) = \psi(v), \quad \forall v \in V, \quad (5.41)$$

with a bounded,  $V$ -elliptic bilinear form  $a(., .)$ . For a discrete bilinear form  $a_h(., .)$  defined on  $V + V_h$ , we define the primal and dual **residual operators**

$$\begin{aligned} \text{Res}(u, v) &= a_h(u, v) - f(v), \\ \text{Res}^*(u^*, v) &= a_h(v, u^*) - \psi(v). \end{aligned} \quad (5.42)$$

**5.2.21 Lemma:** Let  $a_h(., .)$  be a bounded bilinear form on  $V + V_h$  and elliptic on  $V_h$  with norm  $\|\cdot\|_{V_h}$  and constant  $\gamma$ . Then, the error  $u - u_h$  admits the estimate

$$\|u - u_h\|_{V_h} \leq \frac{1}{\gamma} \|\text{Res}(u, .)\|_{V_h^*} + \left(1 + \frac{\|a_h\|}{\gamma}\right) \inf_{w_h \in V_h} \|u - w_h\| \quad (5.43)$$

*Proof.* First, by the definition of the residual, we have the error equation

$$a_h(u - u_h, v_h) = \text{Res}(u, v_h), \quad \forall v_h \in V_h. \quad (5.44)$$

Inserting  $w_h - u_h$  for an arbitrary element  $w_h \in V_h$ , we obtain

$$a_h(w_h - u_h, v_h) = \text{Res}(u, v_h) - a_h(u - w_h, v_h), \quad \forall v_h \in V_h.$$

Using  $v_h = w_h - u_h$  and ellipticity, we obtain

$$\begin{aligned} \gamma \|w_h - u_h\|_{V_h}^2 &\leq a_h(w_h - u_h, w_h - u_h) \\ &= \text{Res}(u, w_h - u_h) - a_h(u - w_h, w_h - u_h) \\ &\leq (\|\text{Res}(u, .)\|_{V_h^*} + \|a_h\| \|u - w_h\|_{V_h}) \|w_h - u_h\|_{V_h}. \end{aligned}$$

Hence, by triangle inequality

$$\|u - u_h\|_{V_h} \leq \frac{1}{\gamma} \|\text{Res}(u, .)\|_{V_h^*} + \left(1 + \frac{\|a_h\|}{\gamma}\right) \inf_{w_h \in V_h} \|u - w_h\|_{V_h}$$

□



**5.2.22 Lemma:** Let  $u \in V$  be the solution to the Poisson equation with right hand side  $f \in L^2(\Omega)$ . Assume  $u \in H^s(\Omega)$  with  $s > 3/2$ . Then, we have for  $v \in V + V_h$ :

$$(f, v) = (\nabla u, \nabla v)_{\mathbb{T}_h} - 2\langle \nabla u, \llbracket v \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} - \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial}. \quad (5.45)$$

*Proof.* We set out from the strong form of the Poisson equation and integrate by parts.

$$(f, v) = (-\Delta u, v) = (\nabla u, \nabla v)_{\mathbb{T}_h} - \sum_{T \in \mathbb{T}_h} \langle \partial_n u, v \rangle_{\partial T}.$$

Under the regularity assumptions of the lemma, all of these integrals make sense at least as duality pairings. In particular,  $\partial_n u \in L^2(\partial T)$ , and thus we can split  $\partial T$  into individual faces. Therefore,

$$\sum_{T \in \mathbb{T}_h} \langle \partial_n u, v \rangle_{\partial T} = 2\langle \nabla u, \llbracket v \otimes \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} + \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial}.$$

The proof concludes by collecting the results.  $\square$

**5.2.23 Lemma:** Let  $k \geq 1$  and let  $V_h$  such that  $\mathbb{P}_{k-1} \subset \Sigma_T$ . Then, if  $u \in H^{k+1}(\Omega)$  and  $v \in V + V_h$ , there holds

$$\begin{aligned} |\text{Res}(u, v)| &\leq ch^k |u|_{k+1} (\|\sqrt{\sigma_h} \llbracket v \rrbracket\|_{\mathbb{F}_h^i} + \|\sqrt{\sigma_h} v\|_{\mathbb{F}_h^\partial}) \\ &\leq ch^k |u|_{k+1} \|v\|_{1,h}. \end{aligned} \quad (5.46)$$

*Proof.* First, we observe that by the regularity assumption,  $\llbracket u \rrbracket = 0$  and thus,  $\mathcal{L}u = 0$ . Hence,

$$a_h(u, v) = (\nabla u, \nabla v)_{\mathbb{T}_h} - (\nabla u, \mathcal{L}v)_{\mathbb{T}_h}.$$

By Lemma 5.2.22 and regularity of  $u$ ,

$$\begin{aligned} \text{Res}(u, v) &= 2\langle \nabla u, \llbracket v \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} + \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial} - (\nabla u, \mathcal{L}v)_{\mathbb{T}_h} \\ &= 2\langle \llbracket \nabla u \rrbracket, \llbracket v \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} + \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial} - (\nabla u, \mathcal{L}v)_{\mathbb{T}_h} \\ &= 2\langle \llbracket \nabla u \rrbracket, \llbracket v \mathbf{n} \rrbracket \rangle_{\mathbb{F}_h^i} + \langle \partial_n u, v \rangle_{\mathbb{F}_h^\partial} - (\Pi_{\Sigma_h} \nabla u, \mathcal{L}v)_{\mathbb{T}_h}, \end{aligned}$$

where  $\Pi_{\Sigma_h}$  is the  $L^2$ -projection. Now, we can apply the definition of the lifting term to obtain

$$\begin{aligned} \text{Res}(u, v) &= 2\left\langle \frac{1}{\sigma_h} \llbracket \nabla u - \Pi_{\Sigma_h} \nabla u \rrbracket, \sigma_h \llbracket v \mathbf{n} \rrbracket \right\rangle_{\mathbb{F}_h^i} \\ &\quad + \left\langle \frac{1}{\sigma_h} (\nabla u - \Pi_{\Sigma_h} \nabla u) \cdot \mathbf{n}, \sigma_h v \right\rangle_{\mathbb{F}_h^\partial}. \end{aligned}$$

Application of standard approximation and trace estimates yields the result observing that  $\sigma_h = \sigma_0/h$ .  $\square$

**5.2.24 Theorem:** Let  $k \geq 1$  and  $V_h$  such that  $\mathbb{P}_k \subset V_T$ . Let  $u \in H^{k+1}(\Omega)$  be the solution to the continuous Poisson problem. Let  $a_h(.,.)$  be the interior penalty method with lifting operators such that  $\nabla V_h \subset \Sigma_h$ . Then, there holds

$$\|u - u_h\|_{1,h} \leq ch^k |u|_{k+1}. \quad (5.47)$$

*Proof.* Application of Lemma 5.2.21, Lemma 5.2.23, and standard interpolation results.  $\square$

**5.2.25 Theorem:** Let the assumptions of Theorem 5.2.24 hold and in addition assume that the problem

$$a(v, u^*) = \psi(v), \quad \forall v \in V,$$

admits the elliptic regularity estimate

$$\|u^*\|_{H^2(\Omega)} \leq c \|\psi\|_{L^2(\Omega)}. \quad (5.48)$$

Then, there holds

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^{k+1} |u|_{H^{k+1}(\Omega)}. \quad (5.49)$$

*Proof.* The proof uses the duality argument by Aubin and Nitsche, which sets out solving the auxiliary problem

$$a(v, u^*) = (u - u_h, v), \quad \forall v \in V.$$

Using the definition of the dual residual, we obtain the equation

$$(u - u_h, v) = a_h(v, u^*) - \text{Res}^*(u^*, v), \quad \forall v \in V + V_h.$$

Testing with  $v = u - u_h$  yields

$$\|u - u_h\|^2 = a_h(u - u_h, u^*) - \text{Res}^*(u^*, u - u_h).$$

Additionally, we use the error equation

$$a_h(u - u_h, v_h) = \text{Res}(u, v_h),$$

tested with  $v_h = I_h u^*$ , to obtain

$$\|u - u_h\|^2 = a_h(u - u_h, u^* - I_h u^*) - \text{Res}^*(u^*, u - u_h) + \text{Res}(u, I_h u^*).$$

Using the regularity of  $u^*$ , the first term on the right admits the estimate

$$|a_h(u - u_h, u^* - I_h u^*)| \leq \|u - u_h\|_{1,h} \|u^* - I_h u^*\|_{1,h} \leq ch \|u - u_h\|_{1,h}.$$

For the second term, we use Lemma 5.2.23 to obtain

$$|\text{Res}^*(u^*, u - u_h)| \leq ch |u^*|_2 \|u - u_h\|_{1,h}.$$

Finally, using  $\llbracket u^* \rrbracket = 0$ , the same lemma yields

$$\begin{aligned} |\text{Res}(u, I_h u^*)| &\leq ch |u|_2 (\|\sqrt{\sigma_h} \llbracket I_h u^* \rrbracket\|_{\mathbb{F}_h^i} + \|\sqrt{\sigma_h} I_h u^*\|_{\mathbb{F}_h^\partial}) \\ &= ch |u|_2 (\|\sqrt{\sigma_h} \llbracket u^* - I_h u^* \rrbracket\|_{\mathbb{F}_h^i} + \|\sqrt{\sigma_h} (u^* - I_h u^*)\|_{\mathbb{F}_h^\partial}) \\ &\leq ch |u|_2 h^k |u^*|_{k+1} \end{aligned}$$

Using the energy estimate in Theorem 5.2.24 we can conclude the prove.  $\square$

# Bibliography

- [Ciarlet, 1978] Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems*. North-Holland.
- [Evans, 1998] Evans, L. C. (1998). *Partial Differential Equations*. American Mathematical Society.
- [Gilbarg and Trudinger, 1998] Gilbarg, D. and Trudinger, N. S. (1998). *Elliptic Partial Differential Equations of Second Order*. Springer.
- [Kondrat'ev, 1967] Kondrat'ev, V. A. (1967). Boundary value problems for elliptic equations in domains with conical or angular points. *Trans. Moscow Math. Soc.*, 16:227–313.
- [Rannacher and Scott, 1982] Rannacher, R. and Scott, L. R. (1982). Some optimal error estimates for piecewise linear finite element approximations. *Math. Comput.*, 38:437–445.
- [Schatz and Wahlbin, 1977] Schatz, A. H. and Wahlbin, L. B. (1977). Maximum norm estimates in the finite element method in plane polygonal domains. Part I. *Math. Comput.*, 32:73–109.
- [Verfürth, 2013] Verfürth, R. (2013). *A Posteriori Error Estimation Techniques for Finite Element Methods*. Oxford University Press.

# Index

- V-elliptic, **17**
- [Gilbarg and Trudinger, 1998, Theorem 8.8], 24
- a posteriori error estimator, **50**
- adjoint problem, **47**
- affine mapping, 37
- approximation error, 60
- Assembling the matrix, 36
- averaging operator, **91**
- Banach fixed-point theorem, 71
- Banach space, **13**
- barycentric coordinates, **30**, 52, 56
- bilinear form, **17**
- bilinear mapping, 37
- boundary, **5**
- Bramble-Hilbert, 41
- broken Sobolev norm, **43**
- bubble function, **56**
- Bunyakovsky-Cauchy-Schwarz inequality, 11, 13, 16
- Céa, 36
- Cauchy sequence, 9, 13
- cell, **27**
- Clément quasi-interpolation, 51
- closed, **13**
- coercive, **17**
- complete, 13, **13**
- completion, **13**
- condition number, *see* spectral condition number
- conforming approximation, **35**
- conjugate gradient method, 75, **77**
- consistency error, 60
- consistent, **86**
- Continuous basis functions, 33
- continuously embedded, **22**
- data oscillation, **55**
- degrees of freedom, **28**
- Dirac  $\delta$ -distribution, **21**
- Dirichlet energy, **6**
- Dirichlet principle, 5
- Dirichlet problem, **5**, 6, 10
- Discontinuous basis functions, 34
- discrete problem, **35**
- distributional derivative, **20**
- domain, **5**, **69**
- dual problem, **47**
- dual space, **16**
- edge, **27**
- efficient, **50**
- elliptic, **17**, **19**, **69**
- Elliptic regularity, 47
- elliptic regularity, **47**, 48, 93, 97
- equivalent, **45**
- error propagation operator, 71
- essential boundary condition, **10**
- face, **27**
- facet, **27**
- finite element, **28**
- finite element space, **28**
- Friedrichs inequality, 8, 9
- Galerkin approximation, 35, **35**
- Galerkin equations, 35
- Global regularity, 25
- gradient, **4**
- Green's function, **48**
- Heaviside function, **20**, 21

Hilbert space, **13**  
 homogeneous, **5**  
  
 inner product, **11**, **11**  
 inner product space, **13**  
 interior penalty method, **92**, **94**  
 Interior regularity, **25**  
 Inverse estimate, **45**  
  
 jump operator, **53**, **91**  
  
 $\kappa(A)$ , **69**  
 $\kappa(\mathbf{A})$ , **70**  
 Kondratev, **25**  
 Krylov space, **77**  
  
 $\Lambda(A)$ , **69**  
 $\lambda(A)$ , **69**  
 $\Lambda(B, A)$ , **75**  
 $\lambda(B, A)$ , **75**  
 $\Lambda(\mathbf{A})$ , **70**  
 $\lambda(\mathbf{A})$ , **70**  
 Laplacian, **4**  
 Lax-Milgram, **18**  
 lifting operator, **94**  
 linear functional, **16**  
 linear mapping, **16**  
 locally quasi-uniform, **50**  
  
 mass matrix, **79**  
 mean value operator, **53**  
 mesh, **27**  
 Method of steepest descent, **75**  
 Meyers-Serrin, **22**  
 minimizing sequence, **9**  
 multigrid iteration, **82**  
  
 natural boundary condition, **10**  
 nodal interpolation, **43**  
 node functional, **28**  
 normal derivative, **5**  
 normal vector, **5**  
 normed dual, **16**  
 normed vector space, **13**  
  
 orthogonal, **13**  
 orthogonal complement, **13**  
  
 orthogonal projection, **15**  
 orthogonal set, **13**  
 orthonormal set, **13**  
  
 Poincaré inequality, **41**  
 Poisson's equation, **4**, **5**  
 pre-Hilbert space, **13**  
 preconditioned cg method, **77**  
 preconditioned Richardson iteration, **74**  
 preconditioner, **74**  
 pull-back, **37**  
  
 quasi-uniform, **43**  
  
 Rayleigh quotient, **70**  
 reference cell, **37**  
 relaxation parameter, **72**  
 reliable, **50**  
 residual, **53**  
 residual operator, **95**  
 Richardson's method, **72**  
 Riesz isomorphism, **16**, **73**, **75**, **77**  
 Riesz representation theorem, **16**, **72**  
  
 Scaling lemma, **39**  
 Schöberl quasi-interpolation, **52**  
 Scott-Zhang quasi-interpolation, **52**  
 shape function, **28**  
 shape function basis, **28**  
 shape regular, **43**  
 Sobolev space, **9**, **21**, **22**  
 spectral condition number, **69**, **70**, **71**,  
     **78**  
 spectral equivalence, **74**, **78**  
 steepest descent, **76**  
 stiffness matrix, **79**  
 Strang's first lemma, **59**  
 stronger norm, **45**  
 symmetric, **69**  
  
 tensor product basis, **31**  
 tensor product polynomials, **31**  
 The Multigrid iteration MGM( $\ell, u_\ell^k, b_\ell$ ),  
     **82**  
 The space  $\mathbb{Q}_2$ , **33**  
 topology, **28**  
 Trace theorem, **23**

unisolvent, **28**

vertex, **27**

weak derivative, **20**

weaker norm, **45**